Applied Games for smart phenotypic data acquisition - challenges of a web platform with digital gamified testing instruments for online-based large scale phenotyping

A Cumulative Dissertation

Submitted to the Faculty of Psychology, University of Basel,

in partial fulfillment of the requirements for the degree of Doctor of Philosophy

by

M Sc Andreas Aeberhard

from Basel (BS), Switzerland

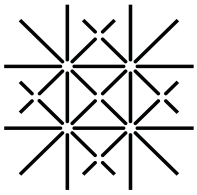
Basel, Switzerland,

November 2019

First-Supervisor: Prof. Dr. med. A. Papassotiropoulos

Second-Supervisor: Prof. Dr. med. Dominique J.-F. de Quervain

Chairperson of the doctoral committee: Prof. Dr. Jens Gaab



University of Basel

Approved by the Faculty of Psychology

At the request of

Professor Dr. med. Andreas Papassotiropoulos

Professor Dr. med. Dominique J.-F. de Quervain

Basel, the _____

_____

Dean

Abstract

Since more than five decades, the replication crisis taints the field of psychological science. Small sample sizes and low statistical power are the identified issues, resulting in a staggering amount of results from studies that can not be replicated. Combating this crisis requires new and scalable approaches that enable innovative testing instruments. Gamification and Applied Games offer those crucial and innovative new ways to assess cognition in an online setting. On the basis of two original research objects, this thesis highlights the challenges, obstacles and benefits from utilizing gamification and applied games for large-scale phenotypic measurements on the basis of an online platform called COSMOS. First, we applied this concept in the form of our established platform COSMOS which we presented in a published paper. In this paper, we purposed a digital psychometric toolkit in the guise of applied games that enables automatized psychometric data collection while measuring a broad range of cognitive functions. Second, we conducted a pilot study that assessed the feasibility and acceptance of a gamified test of the N-back task called HoNk-Back. We showed that participants like the HoNk-Back more than the non-gamified N-back and are more likely to replay the HoNk-Back again. Both the paper and the pilot study point out the benefits of using gamification and applied games for large-scale neurocognitive phenotype cohort screenings for genetic and imaging studies. The challenges and hurdles for future studies regarding data protection, data security and ethics in the online acquisition of personal data are identified and discussed.

Contents

List of Figures

Acknowledgments

I would like to thank both my supervisors Professor Dr. med. Andreas Papassotiropoulos and Professor Dr. med. Dominique J.-F. de Quervain for their continuous support, the opportunity and freedom to conduct research and completing my PhD.

Further, I thank my parents, friends and colleagues for supporting me in so many ways, including listening to my thoughts and my daily challenges that needed to be tackled.

My deepest gratitude and special thanks go toward Dr. Christian Vogler for his tremendous support and tireless assistance. His guidance and the knowledge he shared with me will always be of great importance. I appreciated the plentiful social exchanges with Dr. Tobias Egli, Dr. Leo Gschwind and Georg Brein. I thank Dr. Joe Kossowsky, Dr. Virginie Freytag and Dr. Gediminas Luksys for their helpful input and countless discussions.

My last words go to Lola Liverpool, who kept reminding me how important focus and dedication is and always supported me on so many levels throughout this incredible journey.

Thank you all.

Abbreviations

CSRF	Cross-Site-Request-Forgery
DVG	Digitale-Versorgung-Gesetz
GDPR	General Data Protection Regulation
HB	HoNk-Back
HTML	Hypertext Markup Language
HTTP	Hyper Text Transfer Protocol
HTTPS	Hyper Text Transfer Protocol Secure
MySQL	My Structured Query Language
NB	N-back
PHP	Personal Home Pages / Hypertext Pre-processor
TLS / SSL	Transport Layer Security / Secure Sockets Layer
ToM	Theory of Mind
UIC	Unique Identifier Code
URL	Uniform Resource Locator
WM	Working Memory

1 Introduction

Psychological science suffers from a "crisis of confidence" and there is "doubt among practitioners about the reliability of research findings in the field" (Pashler & Wagenmakers, 2012). This crisis originated in the 1970s and has been going on for more than five decades (Elms, 1975). Recently, several replication studies were unable to replicate results from previously published high-profile psychological studies (e.g. Doyen, Klein, Pichon, & Cleeremans, 2012; Nosek & Lakens, 2014; Pashler, Coburn, & Harris, 2012; Pashler et al., 2012), thus casting doubt on well-established psychological phenomena (Lilienfeld, 2017). Furthermore, it is estimated that more than 50% of the research results are false and therefore irreproducible (Ioannidis, 2005), failing critical scientific scrutiny. In this context, a prominent paper published by Nosek et al. (2015) found that from 100 studies of three top tier psychological journals, a mere 36% were statistically significant in new replication studies. Some authors even state that "the average power of typical psychological research is estimated to be embarrassingly low" (Perugini, Gallucci, & Costantini, 2014), suggesting that studies should be planned more carefully to have proper statistical power by choosing an appropriate sample size (S. F. Anderson, Kelley, & Maxwell, 2017), where "appropriate" often means a larger sample size (Etz & Vandekerckhove, 2016). However, this phenomenon does not only affect psychology, but also other sciences, such as medicine (Begley & Ellis, 2012). It is therefore reasonable to say that new approaches are urgently needed to overcome this crisis, as many authors suggested (e.g. Eich, 2014; Kosara & Haroz, 2018). To

further catalyze this need for modern approaches, Neuroscience, as a rapidly advancing research discipline, requires scalable, efficient and innovative testing instruments. That is why it must be possible for these new approaches to frequently and economically test known parameters and thus obtain large data sets, as this is not possible in standard neuropsychological assessments. With these large data sets, finely calibrated tools can be developed that are sensitive enough to be used in large-scale cohort screenings, diagnostics or assessments.

Part of these novel approaches could be *Gamification*, the use of video game elements in non-gaming systems (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011) created for the purpose of entertainment (Groh, 2012). It is used to improve user experience and user engagement (Deterding, Sicart, et al., 2011), became a widely used technique across various contexts and has been growing rapidly (Landers, 2014). Gamification can be applied to improve participants motivation when doing unattractive tasks and activities (Francisco-Aparicio, Gutiérrez-Vela, Isla-Montes, & Sanchez, 2013) and, in a review by Hamari, Koivisto, Sarsa, et al. (2014), the majority of the reviewed studies showed that gamification yielded positive results and effects on various aspects (e.g. intrinsic vs. extrinsic motivation to complete tasks (Eickhoff, Harris, de Vries, & Srinivasan, 2012), satisfaction (Guin, Baker, Mechling, & Ruyle, 2012) and enjoyment (Mirza-Babaei, Nacke, Gregory, Collins, & Fitzpatrick, 2013)).

Applied Games, on the other hand, are defined as "any form of interactive computer-based game software for one or multiple players to be used on any platform and that

has been developed with the intention to be more than entertainment" (Ritterfeld, Cody, & Vorderer, 2009). They use, on contrast to gamification, gaming as a primary medium (Fleming et al., 2014) rather than just adding game elements to a non-game context as gamification does (Deterding, Dixon, Khaled, & Nacke, 2011) and try to improve the users knowledge, skills, or attitudes (Graafland et al., 2014). Both applied games and gamification try to use games and game-like elements to change patterns of user behavior or experience (Fleming et al., 2017) by focusing on entertainment (Groh, 2012; Winn & Heeter, 2006). According to Deterding (2012), games can leverage both motivation and engagement, one of the common challenges in psychological testing (Gregory, 2004).

Combining the positive effects of increased motivation and engagement in gamification and applied games and its resulting repeated measures might lead to a larger sample size and thus might prevent low statistical power, bringing back reliability, confidence and quality data into psychological science. The use of applied games in an online environment could provide efficient, scalable and innovative testing instruments for measuring cognition. We challenge those statements with our developed research platform *COSMOS* (*CO*gnitive *S*cience *M*etrics *O*nline *S*urvey). There, we implemented the principals of gamification and applied games in the shape of innovative digital gamified testing instruments and brought them together in an online psychometric toolkit for a smart and scalable phenotypic data acquisition with a fully computerized evaluation to measure a wide range of cognitive functions and parame-

ters. This, broken down on a single individual, allows for a relatively low outlay and cost connected to its psychometric testing, further enabling an increased sample size and thus raise statistical power and reliability while being cost-economic. The application example for COSMOS is the easy and cost-effective recruitment and screening of large numbers of subjects for genetic and imaging studies.

This doctoral thesis aims at contributing to the research field of psychology in three ways: firstly, by emphasizing important security aspects on how a web platform has to be conceptualized, built and maintained in order to realize and run online applied games for an automatized and smart phenotypic data acquisition, secondly, by showing that phenotypic measurements are achievable and feasible through online-based tests, thirdly, by discussing and highlighting the resulting challenges and how to tackle them. It includes the following publication:

- Aeberhard, A., Gschwind, L., Kossowsky, J., Luksys, G., Papassotiropoulos, A., de Quervain, D., & Vogler, C. (2018). Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice. *Journal of Technology in Behavioral Science.*

2 Theoretical Background

Data is more valuable than oil, according to various reports (e.g. The Economist, 2017; Wired, 2014). With this statement, *Big Data*, defined by Gartner (2019) as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" is gaining much more importance. Consequently, it is not surprising that Big Data is today's *Digital Oil* (Yi, Liu, Liu, & Jin, 2014). However, for Big Data you need new, modern statistical tools, data management programs and hardware to manage the large amounts of data and uncover its knowledge (Sivarajah, Kamal, Irani, & Weerakkody, 2017), as it is virtually impossible with traditional software and hardware (Frost, 2015). If one overcomes these hurdles and finds suitable ways to analyze Big Data, it will be possible to achieve greater statistical power (Breur, 2016).

Big Data that captures everything in an uncontrolled and unstructured way (for e.g. Facebook, which collects two-digit petabytes (1 petabyte equals 10^{15} bytes or 1'000'000 gigabytes) log data per month (Idrees, Alam, & Agarwal, 2018)) could lack data quality (Taleb, Dssouli, & Serhani, 2015). Wherefore you have an overall advantage if you design a platform from the beginning in such a way that it collects Big Data in a controlled and structured data set, even if the dataset becomes smaller this way. That is why it is even more surprising that there are no science-based digital gamified testing instruments available yet that enable online phenotypic Big Data

acquisition or large-scale psychological assessment. An example for Big Data acquisition in the field of psychology is *Cambridge Analytica*, a company that recently made bad headlines by using vast amounts of Facebook user data to help political campaigning in 2016 (see subsection Data Privacy & Security Aspects). A better, more ethical example of using Big Data could be the screening for large-scale neurocognitive genetics studies with an extreme phenotype study design where it is essential to find extremes as they show a greater genetic effect (Emond et al., 2012).

On the basis of this background, we have identified the current status of the methods used in psychological research as follows: (a) most tools do one-shot measures with small reference cohorts because most tests require a neuropsychologist to be physically present, which makes the test much more expensive and it is therefore difficult to obtain large amounts of data; (b) psychometrics is merely doing small data and is non-digital; (c) psychological assessment is not scalable because it uses an outdated toolkit that is not built on modern technological opportunities; (d) psychological tests and diagnostic tools in applied psychology for predicting working memory performance are currently not evidence-based (e.g. reading span task (Daneman & Carpenter, 1980), operation span task (Turner & Engle, 1989)); (e) user experience has been broadly neglected.

Based on these identified problems, we concluded the following needs: (a) engaging and adaptive tools that can deliver high-resolution and longitudinal data for repeated and continuous data collection in order to obtain high-quality data; (b)

digitalized psychometrics for efficient profiling, automation and Big Data analytics, where the neuropsychologist is required only for the evaluation of the data; (c) performance-based and scalable psychological assessment that harnesses modern technologies possibilities; (d) evidence-based tests and diagnostic tools that can predict working memory performance; (e) an entertaining and enjoyable user experience.

We believe that these needs can be addressed by our basic testing battery hosted on COSMOS which consists of the five games *HoNk-Back*, *Drag Race*, *Frog Life*, *Shortcuts* and *Joyrate* and a performance visualization tool.

With this testing battery, we aim to (a) recruit and screen a large number of subjects for large-scale neurocognitive genetics studies to identify and recruit the extremes of the phenotype distribution with the expectation that they have a stronger genetic effect (Emond et al., 2012); (b) test various components of human cognition as e.g. working memory, attention, impulse control and reaction time which are related to fluid IQ (Colzato, Van Wouwe, Lavender, & Hommel, 2006; Engle, Tuholski, Laughlin, & Conway, 1999; Heitz, Unsworth, & Engle, 2005) with evidence-based tools; (c) assess various aspects of decision making and strategic thinking using different game-based complex decision scenarios; (d) developed the first game-based test that measures Theory of Mind[1] (ToM) in the general adult population; (e) provide

[1]Theory of mind, first used by Premack and Woodruff (1978), is defined as "the cognitive capacity to represent one's own and other persons' mental states, for instance, in terms of thinking, believing, or pretending." (Brüne, 2005), or like Harrington, Siegert, and McClure (2005) simply put it, "It is thinking about thoughts."

an entertaining and enjoyable user experience.

In order to achieve a high level of participation, however, we had to give the user an incentive to participate in COSMOS in order to maintain a certain level of motivation.

2.1 Motivational Aspects

Intrinsic motivation is defined as ”doing something because it is inherently interesting or enjoyable” (Ryan & Deci, 2000). We tried to foster intrinsic motivation for our tools by using the approaches of gamification and applied games to ensure an enjoyable and entertaining game experience. There are plenty of ways to make a game enjoyable and entertaining, from the player being challenged by the game (Schmierbach, Chung, Wu, & Kim, 2014; van den Hoogen, Poels, IJsselsteijn, & de Kort, 2012), to the feeling of being in control (Limperos, Schmierbach, Kegerise, & Dardis, 2011; Trepte & Reinecke, 2011) and an easy to control interface (Browne & Anand, 2012). We have tried to incorporate as many of these aspects as possible into our games, hoping that users play more and often, further leading to repeated measures. Additionally, we try to motivate the public to partake by using not only these entertainment aspects, but also by offering the possibility for custom-built performance feedback provided by a visualization tool, where the user can compare her/his game statistics with that of other users.

With these efforts, we think it will be possible to obtain a larger sample size and

Big Data by repeated measurements and eventually greater statistical power.

2.2 Online Testing

As previously mentioned, Neuroscience requires scalable, efficient and innovative testing instruments. The developed online tools fulfill all three points: they are scalable and can provide high-throughput psychometrics; they are efficient and can set new standards in test economic efficiency by acquiring Big Data; they are innovative as they observe performance and behaviors in virtual world setups instead of relying on self-reports and classical tests. Figuratively speaking, they are virtual labs that operate 24 hours, 7 days a week and accessible worldwide with no lab supervisor needed.

An important part of these online tools are their technical facets. In contrast to classical laboratory tests, all user data is potentially available online for everyone if not properly protected. Potential data thieves, hackers or other kind of cyber criminals are the greatest threat to every website, may it be an online store, a blog or, in our case, a platform for psychometric tools. It is consequently not surprising that cybersecurity was reported to be one of the highest security priorities in about 90% of companies worldwide and that online data theft and cybercrime in general has become a lucrative, illegal business in recent years (McAfee, 2014). The only protection that fends of cyber criminals is a combination of security measurements that are put in place to protect the data. That is why security and data privacy are

two main pillars the COSMOS platform is build upon.

2.3 Data Privacy & Security Aspects

Data privacy and data ownership has become an increasingly important topic in recent years, notably in the field of Big Data (e.g. Sagiroglu & Sinanc, 2013; Tene & Polonetsky, 2011; Terzi, Terzi, & Sagiroglu, 2015; Xu, Jiang, Wang, Yuan, & Ren, 2014). Just recently, in May 2018, the European Union issued a new regulation on data ownership and data protection called *General Data Protection Regulation* (short: *GDPR*), which ensures the protection of personal data within the EU (European Parliament, 2016). Most importantly, this regulation also allows the user to access and, if she/he so requests, delete the personal data that a website collected and stores about her/him. GDPR was an important and necessary step, as a recent and prominent example of insufficient data privacy shows: At Facebook, 50 million user profiles for a total of $1m have been harvested and used by Cambridge Analytica without the users consent to create software that predicts and influences how US voters will vote at the polls (Cadwalladr & Graham-Harrison, 2018). Of these 50 million users, no one has received a monetary share of the $1m paid by Cambridge Analytica. One could therefore argue that the users' data was treated as Facebooks property. This example not only shows that personal data and Big Data is valuable and needs to be properly secured and protected, but also that ethical aspects are involved and play a major role when using or selling personal data. From an ethical

point of view, the user should be entitled to a share of the profit from the sale of her or his data if she or he gives her/his consent to the sale of the data. Furthermore, she/he should be fully informed of what data is being sold and for what purpose it is being used before consent is given. This should also be applied when the users' personal information is used for scientific purposes. However, it can be argued that instead of monetary remuneration, non-monetary remuneration in the form of feedback could be considered, since most academic institutions are rarely able to pay large amounts for the compensation of subjects.

Securing and protecting the digital environment, like e.g. the web server on which the users' personal data is stored, goes hand in hand with data privacy and data security. Achieving high security with a web server can be done in dozens, if not hundreds of ways, from simply choosing an adequate password during the installation to the utmost sophisticated method of encrypting the data (e.g. Deepa and Thilagam (2016); Muscat (2016) or Goseva-Popstojanova, Anastasovski, Dimitrijevikj, Pantev, and Miller (2014) for an overview). Not doing so can have fatal consequences, as seen in the *Yahoo Inc.* incident in 2013, where a staggering 1 billion user accounts and their corresponding data, such as names, email addresses, telephone numbers, dates of birth and hashed passwords were stolen (Trautman & Ormerod, 2016) and sold on a darknet marketplace (Cox, 2016). This incident is no exception (for an exemplary list see `https://en.wikipedia.org/wiki/List_of_data_breaches`). Data stolen in this fashion is predominantly sold on marketplaces on a special network on the

Internet called *Darknet*.

Darknet refers popularly to its own network that supports cryptographically hidden websites that primarily offer criminal services including but not limited to hacking (Moore & Rid, 2016). The sale of such illegal information with no fear of legal consequences is possible through the complete anonymization of the user and the fact that the users in the darknet cannot be tracked due to special technical features (Nunes et al., 2016). Stolen Big Data datasets can be sold quickly, easily, profitably and safely in the darknet with no legal consequences, suggesting that security and therefore data privacy must be a very important aspect in the development of every online environment, especially when collecting Big Data.

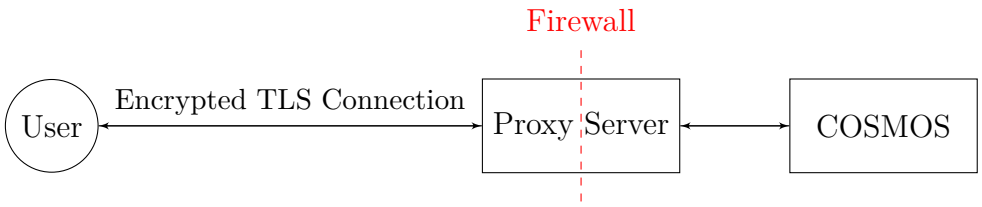


Figure 1. Simplified figure of the server setup with the Proxy in the center. The secure connection to COSMOS runs via the Proxy, which acts as an intermediary and firewall.

In the development of COSMOS, we have taken into account a large number of suggested security aspects, from relatively simple to highly sophisticated ones to ensure data privacy and data security. An example for a sophisticated security mechanism is the use of a *Proxy Server* (short: *Proxy*) as well as a secure data connection. A Proxy functions as an intermediary server and firewall for user requests and does not allow a direct connection to the COSMOS server. Additionally, the secure data

connection ensures that no third-party can view the data exchanged between the user and the Proxy, guaranteeing data privacy and data security. Figure 1 schematically illustrates this setup. A more detailed description of the security and server setup of COSMOS can be found in the Appendix.

3 Original Research

3.1 Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice

Aeberhard, A., Gschwind, L., Kossowsky, J., Luksys, G., Papassotiropoulos, A., de Quervain, D., & Vogler, C. (2018). Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice. *Journal of Technology in Behavioral Science*.

Journal of Technology in Behavioral Science
https://doi.org/10.1007/s41347-018-0071-5

Introducing COSMOS: a Web Platform for Multimodal Game-Based Psychological Assessment Geared Towards Open Science Practice

Andreas Aeberhard[1] · Leo Gschwind[1] · Joe Kossowsky[2,3,4] · Gediminas Luksys[1,5,6] · Andreas Papassotiropoulos[1,7] · Dominique de Quervain[7,8] · Christian Vogler[1,7]

Abstract

We have established the *CO*gnitive *S*cience *M*etrics *O*nline *S*urvey (COSMOS) platform that contains a digital psychometrics toolset in the guise of applied games measuring a wide range of cognitive functions. Here, we are outlining this online research endeavor designed for automatized psychometric data collection and scalable assessment: once set up, the low costs and expenditure associated with individual psychometric testing allow substantially increased study cohorts and thus contribute to enhancing study outcome reliability. We are leveraging gamification of the data acquisition method to make the tests suitable for online administration. By putting a strong focus on entertainment and individually tailored feedback, we aim to maximize subjects' incentives for repeated and continued participation. The objective of measuring repeatedly is obtaining more revealing multitrial average scores and measures from various operationalizations of the same psychological construct instead of relying on single-shot measurements. COSMOS is set up to acquire an automatically and continuously growing dataset that can be used to answer a wide variety of research questions. Following the principles of the open science movement, this data set will also be made accessible to other publicly funded researchers, given that all precautions for individual data protection are fulfilled. We have developed a secure hosting platform and a series of digital gamified testing instruments that can measure theory of mind, attention, working memory, episodic long- and short-term memory, spatial memory, reaction times, eye-hand coordination, impulsivity, humor appreciation, altruism, fairness, strategic thinking, decision-making, and risk-taking behavior. Furthermore, some of the game-based testing instruments also offer the possibility of using classical questionnaire items. A subset of these gamified tests is already implemented in the COSMOS platform, publicly accessible and currently undergoing evaluation and calibration as normative data is being collected. In summary, our approach can be used to accomplish a detailed and reliable psychometric characterization of thousands of individuals to supply various studies with large-scale neurocognitive phenotypes. Our game-based online testing strategy can also guide recruitment for studies as they allow very efficient screening and sample composition. Finally, this setup also allows to evaluate potential cognitive training effects and whether improvements are merely task specific or if generalization effects occur in or even across cognitive domains.

Keywords Gamification · Smart data acquisition · Online phenotyping · Next-generation high-throughput psychometrics

✉ Christian Vogler
christian.vogler@unibas.ch

[1] Department of Psychology, Division of Molecular Neuroscience, University of Basel, Basel, Switzerland

[2] Program in Placebo Studies and the Therapeutic Encounter, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, MA, USA

[3] Department of Anesthesiology, Perioperative and Pain Medicine, Boston Children's Hospital/Harvard Medical School, Boston, MA, USA

[4] Department of Clinical Psychology & Psychotherapy, University of Basel, Basel, Switzerland

[5] Centre for Discovery Brain Sciences, University of Edinburgh, Edinburgh, UK

[6] ZJU-UoE Institute, Zhejiang University School of Medicine, Haining, Zhejiang, China

[7] Psychiatric University Clinics, Basel, Switzerland

[8] Department of Psychology, Division of Cognitive Neuroscience, University of Basel, Basel, Switzerland

Published online: 28 September 2018

Introduction

Objectively measuring inter- and intra-individual differences in human behavior is a fundamental core mission in psychology as it provides the solid fundament on which the entirety of research endeavors in psychology and related fields depend upon (Jenkins and Lykken 1957). The availability of accurate, reliable, and comprehensive phenotypic measures is not only essential for psychological hypothesis testing per se, but is also crucial for the successful elucidation of biological underpinnings of neurocognitive traits that are amenable to for instance imaging or genetic studies (Congdon et al. 2010). While computers have already been used to assist in test evaluations for more than half a century (Kleinmuntz 1963), advances in computer technology now allow for the development of completely digitalized assessment strategies with automated scoring and evaluation procedures (Luciana 2003). Automatization of psychometric assessment is a highly valuable approach for meeting for example the demands that are put forward by the recent revolutions in biotechnology: while high-throughput cost- and time-efficient individual whole genome scans in large cohorts have become a matter of course, phenotypic assessments typically still rely on laborious testing batteries, often requiring trained administrators and stationary attendance time of study participants.

We argue that bringing down the effort for both researchers and testees involved in collecting repeated phenotypic measurements of healthy large cohorts is feasible through online-based test administration. Yet, this requires a substantial redesign and redevelopment of psychometric assessment procedures and instruments. Conceptualizing the novel strategies for large-scale assessments should be led by the idea that participant compensation is essential and constituted by existing ethical guidelines yet does not necessarily need to be monetary. Entertainment that can be achieved through gamification and task design is not only a highly valued benefit itself, it is also key to nurse the participant's motivation required for repeated measurements (Lumsden et al. 2016). Designing the data collection process as a rewarding experience itself is a valuable strategy, as previous studies have found that the demanding nature of data entry is one of the primary reasons respondents stopped using health apps (Krebs and Duncan 2015). Additionally, automation of data collection and evaluation can be used to provide test persons with graphically illustrated feedback on their own performance as this also serves as an incentive for repeated and continuous participation. Finally, computer game-based tests and experiments provide scientists with a novel technique to test ecological validity of laboratory-based procedures, which is always assumed, but rarely tested (Krakauer et al. 2017).

A large online-based research platform that collects sensitive personal data requires continuous attention and efforts to ensure the best possible standard of security for safeguarding participants' personal data from security gaps and potential misuse. It is a question of respect towards the study participants to view and treat the gathered data as a good that the scientist is only entrusted with for conducting research, but that ultimately still belongs to the testees. The fact that the data might be used in currently undefined future research projects or may yield to potential monetization of research outcomes calls for more control options through participants during the data life cycle than a single "open ended" consent form (Lipworth et al. 2017). Yet, despite these concerns, using a single platform framework to simultaneously obtain a wide variety of different psychometric data comes with a set of very appealing options: based on the concepts of "open-science," "open-data," and collaboration, we outline our prototype for automatized and smart phenotypic data acquisition, which holds the potential for reshaping standard procedures in psychological research practice and for facilitating productivity and study outcome reliability. Specifically, we plan to implement a pre-registration system that grants publicly funded scientists' script-based access to the collected data through the COSMOS platform. Scientists can develop their scripts on a dummy database system that mimics the database system of the COSMOS backend. Relying on script-based analyses, which will be run in a secure environment and only return the result of the analysis, is a safety precaution which eliminates the need to grant access to raw data. Only revealing combined and summarized data still allows making highly flexible and efficient use of the existing data pool, while maximizing the security of the dataset against identifying individual test participants.

Conveniently, the ongoing automatic data acquisition continuously generates novel samples that can be used for effortlessly replicating the obtained findings as soon as a large enough additional batch of data has been collected. Additionally, the comparably low maintenance and personnel costs of data gathering can contribute to alleviate the time-consuming competition over limited funding resources. At the same time, the centralization of longitudinal data gathering enables a higher phenotypic resolution per individual than single studies could achieve. The large N high-resolution data allows building models of higher complexity that are better suited to account for confounding factors, which typically would be out of scope for small N single-hypothesis testing study designs. Depending on the respective research question and the hypothesis tested, the available detailed assessment of a large number of individuals allows the application of sampling strategies that either are currently not taken into consideration at all or are only feasible at large expenses of cost and time: preselecting subgroups as homogenous as possible, closely matching experimental groups on potentially confounding factors, evaluating whether a detected correlation can be found in a set of different subgroups, whether it is largely stable or may be even reversed along the continuum of the normal distribution of a given trait.

Finally, platforms like COSMOS can facilitate settling the question, whether so-called brain training (i.e., repeatedly engaging in cognitively demanding tasks) can actually have generalizing beneficial effects: based on a large N, without taking

money from the participants and thus without the inherent conflict of interest the brain training industry-affiliated scientists are faced with.

Game Tests

The COSMOS platform (https://cosmos.psycho.unibas.ch/) is now in its pilot phase, hosting five prototypes of games that currently undergo refinement and calibration as psychometric testing instruments, which are described in more detail below. Table 1 gives an overview of all developed instruments together with the phenotypic constructs they have been designed to measure.

HoNk-Back

The HoNk-Back task is a gamified redesign of one of the most widely used working memory tasks in neuroscience, the N-Back paradigm (Owen et al. 2005). This gamification of the task goes beyond simply adding game-like reinforcement mechanics such as a score or a progress bar. We put special attention on developing a setting that lets the actual task of monitoring a sequence of stimuli appear as natural and plausible as possible, aiming at increasing ecological validity. The task setting makes the test subject to assume the role of a truck driver who gets overtaken by a constant stream of cars. Cars appearing in the review mirror trigger the required response signal by the truck driver which consists of either flashing the headlights at cars that also gave a light signal or waving at the cars that overtook the truck without emitting a headlight signal. Tilting of the rearview mirror controls the N condition as this allows regulating the number of cars disappearing into the blind spot.

Drag Race

This test in form of a drag race game is designed to measure reaction times to unpredictable and predictable cues and variation in response time accuracy. A light signal sequence of two yellow lights indicates that the driver has to get ready. The green light that indicates the take-off signal then is given after a variable random time interval allowing the measurement of spontaneous reaction time (SRT). The process of shifting gears requires a defined motor response pattern: releasing the accelerator button (spacebar), hitting the gear-shifting button (return) and releasing it again, and pushing the accelerator button again. The gear-shifting procedure is used to record the response times to predictable signals: the revmeter continuously moves towards the optimal switching moment, when the response pattern has to be executed. This allows measuring several reaction times of simple motor responses in the form of foreseeable reaction times (FRT). Evaluating repeated runs allows assessing variation in response time accuracy. We are aware of software and/or hardware-related issues concerning reaction time measurements such as monitor response time, operating system design, and input device-related delay such as key debouncing time (Garaizar et al. 2014; Salmon et al. 2017) that impact the accuracy of the response time measurements. Nevertheless, the provided test should yield rough estimates of individual response times and allow group comparisons under the assumption of equally distributed noise. Also the argument has been brought forward that the error introduced by response devices is bound to be small relative to human variability and will only exert potential effects in experiments that lack statistical power in the first place (Damian 2010). Given that the game will be made freely available as a standalone application, it can serve as test instrument in a controlled lab-based environment with identical hard- and software setups allowing unbiased inter-individual comparisons.

Frog Life

Frog Life is a combinatorial task with increasing difficulty levels consisting of a go/no-go paradigm to measure sustained attention and additionally assesses visual vigilance. The task setting lets participants control a virtual frog in a pond that feeds on dragonflies (go-condition) while avoiding devouring hornets (no-go condition). Simultaneously, the testee needs to escape predators, which are announced through changes in coloring of three different display details, namely the color of the water in the pond, the clouds, or a depicted bush (Fig. 1). Insects only become catchable after they entered the proximity range outlined by a spherical contour around the frog. Snatching of the insects is achieved by pressing the corresponding left or right cursor buttons of the keyboard depending on which side of the screen the insects emerged from. Correct responses of the go-task (eating dragonflies) are rewarded with increasing of the score, while incorrect responses to the no-go condition (eating hornets) decreases the score. Color changes of one of the three display details announce an upcoming predator and require the player to trigger an escape jump by pressing the spacebar. Faster reaction times to the color changes are rewarded with more points Yet, pressing the spacebar while no actual color change is taking place causes the player to lose one of three health points indicated by hearts. If all health points are lost, the player character is granted "game-over." After every successful escape, the game mechanics difficulty level is increased. In case the player fails to detect a color change of the display details, appearance of a predator terminates the game. The color change thus constitutes an additional go/no-go task based on signal detection.

Shortcuts

This game is designed as a two-tier short-term memory performance assessment consisting of an episodic picture recognition

Table 1 Overview of all games that have been developed for the COSMOS platform to date. The column "Features" shortly describes game setup and content. Also listed are approximate durations to complete one testing unit (depending on the game type, the test unit refers to a single level or to a complete game). "Technological specifications" lists the IT technologies/programming languages that were used to create the game and its backend

Game name	Target phenotypes	Features	Implemented	Approx. duration per test unit	Technological specifications
BANKWORLD	Risk-taking behavior, decision-making, strategic thinking	Text-based; managing financial projects that requires social investing to minimize risk of averse events	Yes	120–360 s	HTML5, CSS3, RestAPI, MySQL
BRAINLAB	Spatial memory	3D, first-person perspective, free roaming; solve a series of tasks while staying oriented in a multileveled maze	No	120–600 s	Unity, MySQL
CAKE	Altruism, fairness, trust, suspiciousness	Text-based; complex social interaction game based on advanced ultimatum game scenarios	No	60–120 s	HTML5, CSS3, MySQL
DRAGRACE	Reaction times, eye-hand coordination	2D; cartoon-styled drag race game	Yes	15–80 s	Unity, MySQL
FROGLIFE	Attention, reaction times, impulsivity, eye-hand coordination	2D; keep a frog alive by escaping predators while feeding on dragonflies and avoiding swallowing hornets	Yes	60–180 s	Unity, MySQL
HONK-BACK	Attention, working memory	3D, first-person perspective; react differently (honk or wave) based on whether an overtaking car flashed its headlight	Yes	30–180 s	HTML5, CSS3, JavaScript, MySQL
INVESTMENT BANKING	Decision-making, risk-taking behavior, multitasking	2D; optimize revenue achieved by a team of bankers through individually rewarding while avoiding reward over-saturation	No	300–600 s	HTML5, CSS3, JavaScript, MySQL
JOYRATE	Theory of mind, humor appreciation	Text-based; guess humor appreciation of others based on how they rated statements, covering different topics like politics, personality, etc.	Yes	120–300 s	HTML5, CSS3, xAPI, MySQL
MEMORY RACER	Episodic long- and short-term memory, eye-hand coordination	3D, racing game; rate and memorize pictorial or text-based stimuli presented during gameplay	No	60–120 s	Unity, MySQL
SHORTCUTS	Episodic long- and short-term memory	3D, panda bear climbing mountains; player needs to use recognition memory for time-saving shortcuts in the climbing route	Yes	90–300 s	Unity, MySQL

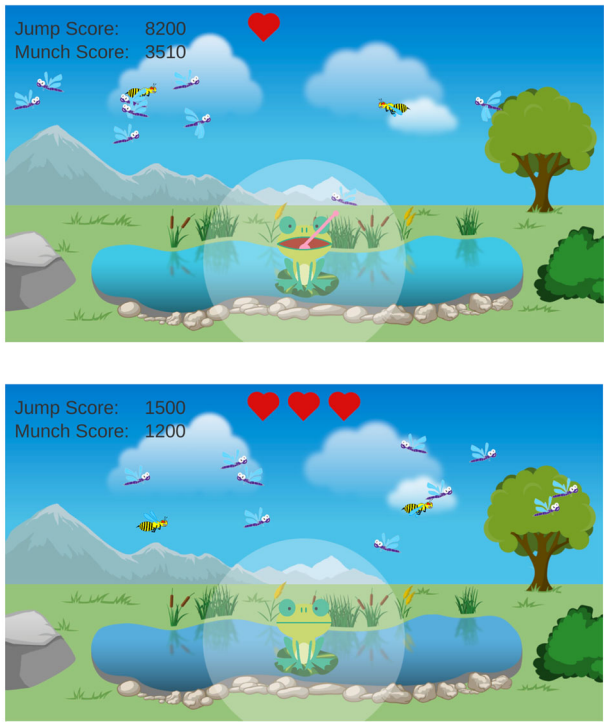


Fig. 1 Scenery examples from the game "Froglife." Dragonflies and hornets constitute a go/no-go paradigm. Eating dragonflies increases the "Munch Score," eating hornets decreases it: upon entering the white circle that is surrounding the frog, the player can make the frog eat the insects using the left and right cursor button, depending on the side from which the insects are entering the white circle. **a** The dragonfly entered from the right side is captured by pressing the right arrow cursor button. **b** Color changes of the pond, the bush on the right side or the clouds indicate an approaching predator requiring the player to escape the current scenery by jumping to a different pond using the spacebar. In the depicted scene, the hue of the pond is changed

task and a sequence-learning test. At the beginning of every game round, testees need to memorize a set of picture stimuli (3 to 10 items). The test participants can choose between different categories, such as "food," "animals," and "sport," and will be presented a set of pictures to memorize for the current game round. Additionally (apart from the "easy" condition featuring only three pictorial items to be remembered), a sequence of differently colored and shaped symbols is presented at the beginning of the game round. The accurate encoding of sequential information is a key cognitive element in human cognition, setting humans apart from other species, but also shows large variation in performance within species (Ghirlanda et al. 2017). During the actual game, the player controls a panda bear climbing rock walls that gets rewarded for correctly solving recognition tasks: at given intervals, the player is presented with a selection of pictures and required to identify the picture shown in the beginning. If she clicks on the correct picture, a bird lifts the panda bear to a higher position in the climbing wall thus rewarding the player with "shortcuts" in the climbing route. Also, at predefined intervals, buttons appear on the screen that require the player to reproduce the symbol sequence that was shown at the beginning of the level. It must be entered correctly in order for the climbing to continue. If the sequence is not entered correctly by the player, the entire sequence will be displayed, so that the player can continue. The player is awarded with points for correct recognition of the pictures, reproducing the symbol sequence correctly and for speed.

Joyrate

This task is primarily designed to measure a subtype of theory of mind (ToM) by employing entertaining stimulus material. At the beginning of the game, the player is asked to rate the jocularity of 10 items consisting of cartoons, memes, and written jokes on a scale from 0 to 10. Additionally, the participants also rate how strongly they agree with 18 statements touching topics such as politics, religion, society, sports, education, and personality. This initial phase has to be completed only once. The actual game then consists in guessing as how amusing a given item has been rated by another person whose ratings the player is randomly assigned to. Whenever an item appears from the joke or the statement pool that has not yet been answered by the player himself, he is asked to make his own rating prior to estimating/learning the estimation of his/her counterpart. This way, the pool of rated jokes and statements for all individuals is constantly increased. The goal of the game is to estimate as accurately as possible how entertaining a given stimuli was perceived by the other person. Apart from the demographic info on the other player that is always provided (gender, age, education), the participant can unlock further information on how the statements were rated using an in-game-generated currency (JokeCoins). The accuracy of the estimation process is rewarded with points and JokeCoins.

COSMOS Environment

Individual Data Visualization

All gamified testing instruments developed in the scope of the COSMOS project feature an application-specific relational SQL database that records the user's input. This makes it easy to set up any application as a standalone implementation and to integrate the applications into a specific laboratory test setting, for example, as a subtest in a given brain mapping experiment. In the scope of the COSMOS web platform, all user data is assigned to unique identifier codes (UIC) and thus to a certain person, by means of a separate central authentication system implemented in the secure software framework used to host the website. All single SQL databases are linked via experience APIs to a Learning Record System employing a mongoDB that serves the purpose of graphically representing the obtained data. We have developed a data visualization application that allows platform administrators fast and easy creation and configuration

of interactive plots. COSMOS participants can choose from a variety of preconfigured plots to learn about their individual performance over time, compare their scores to all participants or to specific subgroups only, e.g., a given age range or gender (examples of plots are depicted in Fig. 2). Visualization of achieved high scores and selected performance measures like for instance average reaction times also allows COSMOS participants to monitor their performance over the course of the day to identify peak performance time periods when they usually achieve the best concentration and attention levels.

Automated Data Processing Pipeline

The independent SQL databases that all games running on the COSMOS platform are equipped with facilitate a streamlined and automated data analysis pipeline. While there may be specific deviations for single games, the general rule is that data will be marked as an unfinished run or simply not stored in the database, if the level was aborted due to player inactivity, closing of the browser, or loss of internet connection. All user responses and summary statistics generated by the games are recorded and stored along with a timestamp and linked to a specific UIC in the games' databases. The UIC is generated when an account is registered and thus pertains to specific login credentials. This procedure allows data to be uniquely assigned to a specific person and therefore enables data collection over multiple trials, time-points, levels, and different tasks. Since the exact timestamp of each reaction is always stored in the database, it is easy to calculate for example the average reaction time per game round: large intervals between stimulus presentation and the reaction of the player or a large variance in task performance indicators can be used to detect a lack of concentration or distraction and thus can be used to create QC filters. Of course, those statistical filters

Fig. 2 Example of a typical visualization of test results generated by the mongoDB-based visualization feature of the COSMOS web platform. The generated graphical representations are partially configurable and allow the user to customize which data is displayed. Any data fed into the mongoDB can be visualized in either bar charts, pie charts, or progress charts. Database schema of the game Frog Life

themselves can be evaluated if participants are asked to rate for example the attention or level of concentration they were exhibiting during gameplay after a level is completed.

The use of standard SQL databases allows accessing the data with all common statistical analysis tools/languages like R, python, matlab, octave (Eaton et al. 2014; MATLAB Optimization toolbox 2017; R Core Team 2018), etc. This allows the creation of standard query scripts that are customizable to retrieve the data best suited to answer a given research question: e.g., retrieve all data for game x, y, and z for all individuals meeting a given age range, gender, or educational level that finished at least 10 trials per game within a specified time period. The exact procedure of reading, processing, summarizing, and blending data may of course depend on the specifics of the research question to be answered.

Figure 3 depicts a description of the SQL database schema for the game Frog Life. This description together with the information on the different response types (as shown in Table 2) helps understanding how simple database queries can be used to sum up different correct answers and/or errors depending on the difficulty level of the task in order to serve as a data basis for modeling.

Modeling Phenotypes

In order to understand and analyze complex behaviors, a promising approach has been computational models (Corrado and Doya 2007; Luksys and Sandi 2011; Mars et al. 2012; Nassar and Frank 2016). Most widely popularized in the field of reinforcement learning (Tanaka et al. 2004; Daw

Fig. 3 The SQL database schema for Frog Life. The games-table records all the games (numbered incrementally starting from 1) a given user (user_id defined by the UIC) has played, along with the scores s/he achieved and the timestamp the game was started (creation_time) and finished (modification_time). The finished field contains the info whether the game was normally finished or prematurely terminated. All lines between the tables are dotted, since the UIC serves as foreign key for all other tables. The rounds-table contains information about every single round played as indicated by the "one to one or many" relationship (since many rounds per game are possible). A round starts either directly at the beginning of the game or after the player escaped an upcoming predator and jumped to a new scenery. After every round, the difficulty level is increased (and the current difficulty level gets stored in the difficulty field), i.e., the speed of the insects accelerates, the color change time decreases, and the hue intensity change gets less pronounced. The action table stores all the actions that are exhibited by the player during a given round. The action_types-table comprises all the possible response types a player can display (see Table 1 for action type definitions). The levels-table holds the information about the background sceneries, which is recorded in the rounds-table (level_id). Currently, three different sceneries are available

Table 2 Description of possible action types a user can display during playing Frog Life. Only if a given response as defined by an action type is exhibited, one of the described database entries in the Description column is triggered. Thus, e.g., if no entry for action type HORNET_EATEN exists in a given round table, the player did not make this type of mistake during that round. Type describes the psychometric characteristics attached to the potential user responses

Action	Description	Type
DRAGONFLY_EATEN	Time difference between required action trigger and correct response	Correct response
DRAGONFLY_NOT_EATEN	Time span of omission error	Omission error/go-error
HORNET_EATEN	Time difference between no-go trigger and incorrect response	No-go error
HORNET_NOT_EATEN	Time span of correct omission	Correct response
WRONG_LEFT_PRESS	Timestamp of pressing the opposite of the required arrow key	Motor control error
WRONG_RIGHT_PRESS	Timestamp of pressing the opposite of the required arrow key	Motor control error
COLOR_REACTION	Time difference between upcoming predator warning and pressing the spacebar (finish current round)	Correct response
COLOR_NO_REACTION	Time difference between upcoming predator and game over	Visual attention error
CAUSELESS_JUMP	Timestamp of pressing spacebar without predator approaching	Visual threshold error
ARROW_NO_INSECT	Timestamp of pressing an arrow button in vain	Motor control error

et al. 2006; Behrens et al. 2007; Frank et al. 2007; Luksys et al. 2009), they have also been applied to study working (Collins and Frank 2012; Collins et al. 2014) and episodic (Luksys et al. 2014, 2015) memory as well as decision-making (Forstmann et al. 2008), including strategic reasoning (Zhu et al. 2012; Seo et al. 2014). The main principle is that a computational model is fitted to experimental data (based on how well model-produced behaviors match experimentally observed ones), and then the best-fitting model parameters and/or variables are used as correlates for neurobiological data such as neuron recordings (Samejima et al. 2005), fMRI activations (Tanaka et al. 2016; Daw et al. 2006, Behrens et al. 2007), genetic differences (Frank et al. 2007; Set et al. 2014; Luksys et al. 2014, 2015), levels of stress (Luksys et al. 2009), and neuropsychiatric disorders (Collins et al. 2014). The main advantage of model-based analysis is that it can test neurocomputational mechanisms of behavior, which different candidate models aim to represent, and reduce a variety of behavioral measures, which can strongly depend on the specific task, to fewer model parameters that are directly comparable between the tasks. For example, reinforcement learning can model behavior in a number of tasks where rewards or punishments of some kind (sometimes implicit) are involved, and despite different formalizations, most of these models have common parameters such as the learning rate, exploration-exploitation tradeoff, and future discounting (Tanaka et al. 2004; Frank et al. 2007; Schweighofer et al. 2008; Luksys et al. 2009). Due to unusual richness of the acquired data, gamification provides a special opportunity to convincingly show usefulness of computational models compared to traditional analyses of behavior, and most importantly link platform-derived behaviors to laboratory-based tasks, which can be analyzed using more simple models that share parameters with more complex models of games. Where explicit modeling of games is not practical (e.g., due to their complexity), the recorded patterns of game-derived data could be linked to laboratory-based behaviors (or their model parameters) using machine learning tools. Finally, a game-based psychometric assessment platform such as COSMOS provides a unique chance to test and compare different candidate models using a much wider variety of tasks and populations than used in most model-based analysis studies, where usually a narrow range of models are tested against each other based on one or few tasks selected by authors (which may benefit their favorite models compared to alternatives).

Discussion

The pervasive problem of low-powered studies in the behavioral and social sciences leading to non-replicable and spurious findings has already been identified more than 60 years ago. Yet today, it does not only still persist, but is even being exacerbated by system design faults (Ioannidis 2015; Smaldino and McElreath 2016; Szucs and Ioannidis 2017): using the amount of published original research as a quality criterion for awarding funding or tenured positions incentivizes increasing the number of publications. This creates a conflict of interest with the researcher's intrinsic goal to maximize study outcome reliability. In addition, the novelty of findings based on a small number of observations is often valued higher than replication in large cohorts (Higginson and Munafo 2016; Nosek et al. 2015; Vinkers et al. 2015). In combination with short-term contracts for the junior scientific staff (Kreeger 2004; Langenberg 2001) that render planning and implementing of larger-scale projects almost impossible as they require substantially more time than single-hypothesis small N studies, the scientific community has formed an

optimal hotbed for keeping the well-known problems alive and prospering.

The ongoing replication crisis of psychological research presses us to figure out how the above-mentioned systemic shortcomings can be overcome. Luckily, computerization and automatization in combination with interdisciplinary cooperation and an open-data philosophy offer a solution to a very basic but crucial problem: in our eyes, sample size is the elephant in the room for improvement of psychological research that needs to be addressed promptly. Towards this end, we have initiated the COSMOS platform: we are striving to facilitate recruitment of study participants through automatization, i.e., creating experimental setups that no longer require staff to implement them and to observe and record behavior. Although our platform is still in its pilot phase, we argue that digitally oriented research endeavors like COSMOS will eventually serve the scientific community in several ways. Online screening platforms can be used to either carefully preselect individuals or to simply increase sample size without skyrocketing costs. Being able to substantially increase the number of study participants is arguably a compelling strategy to counteract the overestimation of effect sizes and the non-replicability of study findings.

For the scaling of psychometric assessment, especially for the online-based test setting, our overall philosophy is that the testing instruments need to be as fun and absorbing for the participants as possible to increase the intrinsic motivation to engage. At the same time, tests should require minimal effort with regard to manual data entry in order to prevent significant issues with subject adherence. Finally, the novel assessment tools should provide investigators in the psychological and biomedical sciences with research-grade cognitive and psychological metrics. Technological advances along with a strongly grown computer literacy in the general population and widespread familiarity with computer games (Granello and Wheaton 2004; Palaus et al. 2017) open up a plethora of possibilities for the operationalization of psychological research questions. Leveraging gamification to repeatedly obtain behavioral samples paves the way for a next-generation high-throughput psychometric toolset. Hence, the COSMOS platform is conceptualized to collect a vast array of psychometric and cognitive data from a large pool of study participants in a highly automated and thus very cost- and time-efficient way.

It is obvious that the goal of gathering in-depth phenotypic data by employing web-based administration of psychometric tests in the guise of entertaining serious games chaperoned by individual automatic performance feedback requires a highly interdisciplinary skill set: social, computer, and data scientists need to work closely together to design, develop, refine, and validate the tools and put them to work. Yet, this aggregate competence is often readily available in university settings and easily accessible through close collaborations between disciplines.

The possibilities of the outlined web platform go way beyond the scope of only gathering data, if additional opportunities offered by the digital era are harnessed: it could also provide a framework to present, discuss, and continuously update scientific findings. We think that eventually such approaches will not only help online participants better understand their own behavior and detect patterns that may be early signs of neuropsychiatric disorders; they could also open up venues for the development of efficient, individualized, and most importantly scientifically sound methods of cognitive enhancement.

Acknowledgments COSMOS (https://cosmos.psycho.unibas.ch/) was started by the Divisions of Molecular and Cognitive Neuroscience of the University of Basel in close cooperation with the Informatics Department of the University of Applied Sciences Northwest-Switzerland.

References

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*(9), 1214–1221.

Corrado, G., & Doya, K. (2007). Understanding neural coding through the model-based analysis of decision making. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 27*(31), 8178–8180. https://doi.org/10.1523/JNEUROSCI.1590-07.2007.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European Journal of Neuroscience, 35*, 1024–1035.

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *The Journal of Neuroscience, 34*, 13747–13756.

Congdon, E., Poldrack, R. A., & Freimer, N. B. (2010). Neurocognitive phenotypes and genetic dissection of disorders of brain and behavior. *Neuron, 68*, 218–230. https://doi.org/10.1016/j.neuron.2010.10.007.

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods, 42*(1), 205–211. https://doi.org/10.3758/BRM.42.1.205.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876–879. https://doi.org/10.1038/nature04766.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007a). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences, 104*(41), 16311–16316. https://doi.org/10.1073/pnas.0706111104.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. Proceedings of the National Academy of Sciences, 105(45), 17538–17542. https://doi.org/10.1073/pnas.0805903105.

John W. Eaton David Bateman, S. H., & Wehbring, R. (2014). {GNU Octave} version 3.8.1 manual: A high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. Retrieved from http://www.gnu.org/software/octave/doc/interpreter.

Garaizar, P., Vadillo, M. A., López-De-Ipiña, D., & Matute, H. (2014). Measuring software timing errors in the presentation of visual stimuli in cognitive neuroscience experiments. *PLoS One, 9*(1), e85108. https://doi.org/10.1371/journal.pone.0085108.

Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: A divide between humans and other animals? *Royal Society Open Science, 4*(6). https://doi.org/10.1098/rsos.161011.

Granello, D., & Wheaton, J. (2004). Online data collection: Strategies for research. *Journal of Counseling & Development, 82*(4), 387–393. https://doi.org/10.1002/j.1556-6678.2004.tb00325.x.

Higginson, A. D., & Munafo, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology, 14*(11), e2000995. https://doi.org/10.1371/journal.pbio.2000995.

Ioannidis, B. J. P. A., & Sc, D. (2015). Failure to replicate: Sound the alarm. Cerebrum, 2015 (November), 1–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27420921. Accessed 21 June 2017.

Jenkins, J. J., & Lykken, D. T. (1957). Individual differences. *Annual Review of Psychology, 8*(1), 79–112. https://doi.org/10.1146/annurev.ps.08.020157.000455.

Kleinmuntz, B. (1963). Personality test interpretation by digital computer. *Science, 139*(3553), 416–418. https://doi.org/10.1126/science.139.3553.416.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., Maciver, M. A., & Poeppel, D. (2017). Neuron perspective neuroscience needs behavior: Correcting a reductionist bias. *Neuron, 93*, 480–490. https://doi.org/10.1016/j.neuron.2016.12.041.

Krebs, P., & Duncan, D. T. (2015). Health App Use Among US Mobile Phone Owners: A National Survey. JMIR MHealth and UHealth, 3(4), e101. https://doi.org/10.2196/mhealth.4924.

Kreeger, K. (2004). Short-term limbo. *Nature, 427*(6976), 760–761.

Langenberg, H. (2001). Uncertainty of short-term contracts is turning talent away from science. *Nature, 410*(6830), 849–850.

Lipworth, W., Mason, P. H., & Kerridge, I. (2017). Ethics and epistemology of big data. *Journal of Bioethical Inquiry, 14*(4), 485–488. https://doi.org/10.1007/s11673-017-9815-8.

Luciana, M. (2003). Practitioner review: Computerized assessment of neuropsychological function in children: Clinical and research applications of the Cambridge Neuropsychological Testing Automated Battery (CANTAB). *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 44*(5), 649–663.

Luksys, G., Gerstner, W., & Sandi, C. (2009). Stress, genotype and norepinephrine in the prediction of mouse behavior using reinforcement learning. *Nature Neuroscience, 12*(9), 1180–1186. https://doi.org/10.1038/nn.2374.

Luksys, G., & Sandi, C. (2011). Neural mechanisms and computations underlying stress effects on learning and memory. *Current Opinion in Neurobiology*. https://doi.org/10.1016/j.conb.2011.03.003.

Luksys, G., Ackermann, S., Coynel, D., Fastenrath, M., Gschwind, L., Heck, A., de Quervain, D. (2014). BAIAP2 is related to emotional modulation of human memory strength. *PloS One, 9*(1), e83707. https://doi.org/10.1371/journal.pone.0083707.

Luksys, G., Fastenrath, M., Coynel, D., Freytag, V., Gschwind, L., Heck, A., de Quervain, D. J.-F. (2015). Computational dissection of human episodic memory reveals mental process-specific genetic. Proceedings of the National Academy of Sciences of the United States of America.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR Serious Games, 4*(2), e11. https://doi.org/10.2196/games.5888.

Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. S. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology, 65*(2), 252–267. https://doi.org/10.1080/17470211003668272.

MATLAB Optimization toolbox. (2017). Retrieved from https://ch.mathworks.com/products/matlab.html.

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*. https://doi.org/10.1016/j.cobeha.2016.04.003.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*, 46–59. https://doi.org/10.1002/hbm.20131.

Palaus, M., Marron, E. M., Viejo-Sobera, R., & Redolar-Ripoll, D. (2017). Neural basis of video gaming: A systematic review. *Frontiers in Human Neuroscience, 11*, 248. https://doi.org/10.3389/fnhum.2017.00248.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/.

Salmon, J. P., Jones, S. A. H., Wright, C. P., Butler, B. C., Klein, R. M., & Eskes, G. A. (2017). Methods for validating chronometry of computerized tests. *Journal of Clinical and Experimental Neuropsychology, 39*(2), 190–210. https://doi.org/10.1080/13803395.2016.1215411.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Neuroscience: Representation of actionspecific reward values in the striatum. *Science, 310*(5752), 1337–1340. https://doi.org/10.1126/science.1115270.

Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S. C., Yamawaki, S., & Doya, K. (2008). Low-Serotonin Levels Increase Delayed Reward Discounting in Humans. *Journal of Neuroscience, 28*(17), 4528–4532. https://doi.org/10.1523/JNEUROSCI.4982-07.2008.

Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science, 346*(6207), 340–343. https://doi.org/10.1126/science.1256254.

Set, E., Saez, I., Zhu, L., Houser, D. E., Myung, N., Zhong, S., et al. (2014). Dissociable contribution of prefrontal and striatal dopaminergic genes to learning in economic games. *Proceedings of the National Academy of Sciences of the United States of America, 111*(26), 9615–9620. https://doi.org/10.1073/pnas.1316259111.

Smaldino, P. E., & McElreath, R. (2016). The Natural Selection of Bad Science. *Royal Society Open Science, 3*(9). https://doi.org/10.1098/rsos.160384.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), 1–18. https://doi.org/10.1371/journal.pbio.2000797.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience, 7*, 887. https://doi.org/10.1038/nn1279.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2016). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. In Behavioral Economics of Preferences, Choices, and Happiness (Vol. 7, pp. 593–616). https://doi.org/10.1007/978-4-431-55402-8_22.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *BMJ, 351*, h6467. https://doi.org/10.1136/bmj.h6467.

Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences, 109*(5), 1419–1424. https://doi.org/10.1073/pnas.1116783109.

3.2 COSMOS Pilot Study

Since its introduction by Kirchner almost 60 years ago (Kirchner, 1958), the N-back paradigm has developed into one of the most commonly used testing concepts to measure working memory (WM) performance. Especially in neuroimaging studies, N-back tasks have found widespread application and emerged as a prototypical measure for WM (Owen, McMillan, Laird, & Bullmore, 2005). The preference for the N-back task in cognitive neuroscience despite other prominent ways to operationalize WM assessment is owed to the fact that test administration is largely unimpeded by methodological constraints given by neuroimaging setups concerning e.g. stimulus-response timing or the available response formats (Redick & Lindsey, 2013). Additionally, the N-back task comes by design with the appealing property that WM load can be parametrically modulated: As participants are presented with a stream of stimuli, they are tasked with deciding whether the currently presented one matches the one presented N items before (Jaeggi, Buschkuehl, Perrig, & Meier, 2010). Since its initial introduction, computerization of the N-back task has led to the development of a series of different variants employing not only auditory but also visual or visuospatial stimuli or combinations of both (Jaeggi et al., 2007; Kidd & Humes, 2015). Typically, psychological test construction is guided by deliberations of operationalization without making special efforts to put the enjoyability for the test subjects into the main focus. Yet, this aspect deserves peculiar attention, if the test is e.g. administered online where disengaging tasks lead to higher dropout rates or if an

investigation targets a given age group like children. Task gamification is a possible solution to the problem as it comes with potentially positive effects on participant engagement and performance, increased participant motivation and test intuitiveness and easily allows augmenting ecological validity by employing comprehensible everyday scenery instead of highly abstract task settings (Burgess et al., 2006; Lumsden, Edwards, Lawrence, Coyle, & Munafò, 2016; Ninaus et al., 2015).

Given that the N-back task could be received as effortful and at times dreary, we set out to expand the available N-back based test armamentarium by a gamified variation of the task we termed *HoNk-Back*. The task setting in HoNk-Back lets the test subject assume the role of a truck driver who gets overtaken by a constant stream of cars. Cars appearing in the wing mirror trigger the required response signal by the truck driver which consists in either flashing the headlights at cars that also gave a light signal or waving at the cars that overtook the truck without emitting a headlight signal. Tilting of the wing mirror controls the N condition as this allows regulating the number of cars disappearing into the blind spot.

Our goal was to develop a test instrument that has a high ecological validity and motivates and involves the participants more. Further, we tried to create a test instrument with a comprehensible everyday scenery instead of highly abstract task settings.

In this pilot study we evaluate the HoNk-Back on the basis of the COSMOS platform. The goal of this study was to evaluate (a) the reproducibility of the classic

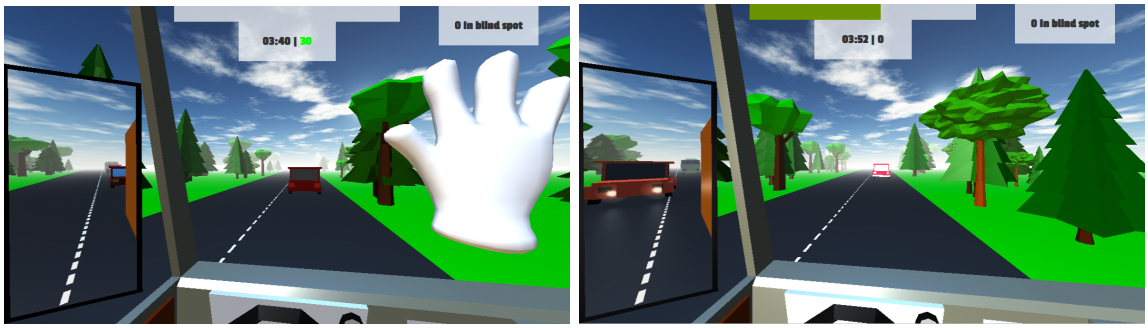
N-back with the HoNk-Back; (b) the comparability of the results; (c) the acceptance of the HoNk-Back; (d) whether the HoNk-Back is more exciting and therefore more motivating than the N-back and (e) the overall feasibility of a gamified test instrument.

3.2.1 Methods

The experiment was administered through the COSMOS platform which consisted of an implementation of a standard N-back task and its gamified version, the HoNk-Back, followed by two additional questionnaires after.

In the HoNk-Back, participants slip into the role of a truck driver who gets constantly overtaken by cars. The cars appear in the wing mirror and overtake the truck and, in some cases, flash their lights while doing so. If a car flashes its lights, the participant needs to flash back as soon as the car overtook him. In the other case where a car does not flash its lights, a hand wave was required. The car's color and shape equaled the different letters of the N-back whereas the tilt angle of the wing mirror of the truck controls the N condition. See figure 2 for illustrating screenshots.

An online adaptation for the standard N-back task was developed. The showtime for the letters in the N-back was set to 500 milliseconds, while the time between letters was set to 1350 milliseconds and the amount of different uppercase letters shown during a level was set to 8, resulting in a total of 116 targets within the 240 seconds (4 minutes). Responses for the N-back could be entered via mouse-click



(a) (b)

Figure 2. HoNk-Back screenshots where the participant waves (a) or flashes the lights (b). Cars overtake on the left side of the truck. Before they overtake, they may or may not flash their lights and disappear in the wing mirror. If they flashed their lights, the participant needs to flash the lights, too (b). If not, they need to wave their hand (a). The tilt of the wing mirror defines the N-back condition: the flatter the tilt angle, the more cars are in the blind spot. Shape and color of the cars equal different letters of the N-back.

or keyboard button-press (see figure 3 for an illustrating screenshot). HoNk-Back responses required keyboard button-presses only.

After completion, participants were redirected to the questionnaire section including demographics, rating the appeal of and readiness to repeat the tasks, weekly hours of engaging in computer games, weekly driving hours and an open-feedback field. At the end, the participants could choose whether they wanted to have automatically generated feedback on their performance, which would allow them to compare themselves with the previous participants of the experiment. Duration of the complete test was roughly 15 minutes. See figure 8 in the appendix for a schematic test arrangement.

We calculated the accuracy of each participant by using the dPrime algorithm with the equation

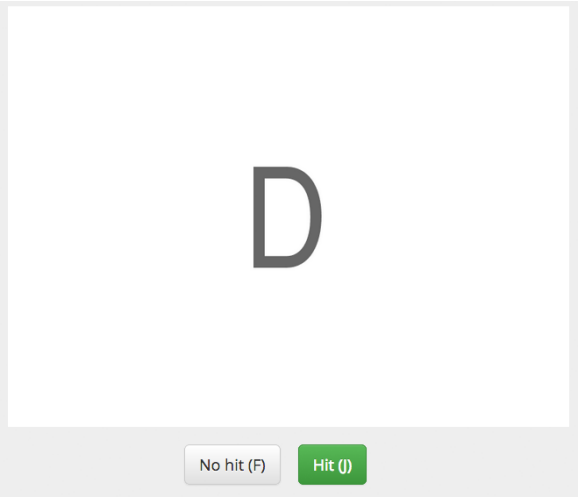


Figure 3. The online adaptation of the N-back was identical to a N-back used in the laboratory. Showtime for the letters was set to 500 milliseconds and 1350 milliseconds for the blank time between letters. In 240 seconds (4 minutes) a total of 116 targets were presented in the format of capital letters. Participants could either enter their responses by clicking on two buttons (*No hit* and *Hit*) or by keyboard-button press (F and J for a no hit or hit, respectively).

$$\frac{(tp + tn)}{(tp + tn + fp + fn + m)} = P \tag{1}$$

where P = performance/accuracy, tp = true positives, tn = true negatives, fp = false positives, fn = false negatives and m = missings, so that $P \in [0, 1]$, where a P value of 1 represents a perfect 100% accuracy and a value of 0 represents 0% accuracy in the respective game.

3.2.2 Participants

A total of N=321 completed the experiment. 37 duplicates were found and removed, resulting in N=284 individuals (191 women, 88 men, 2 rather not say, 3 missings) used

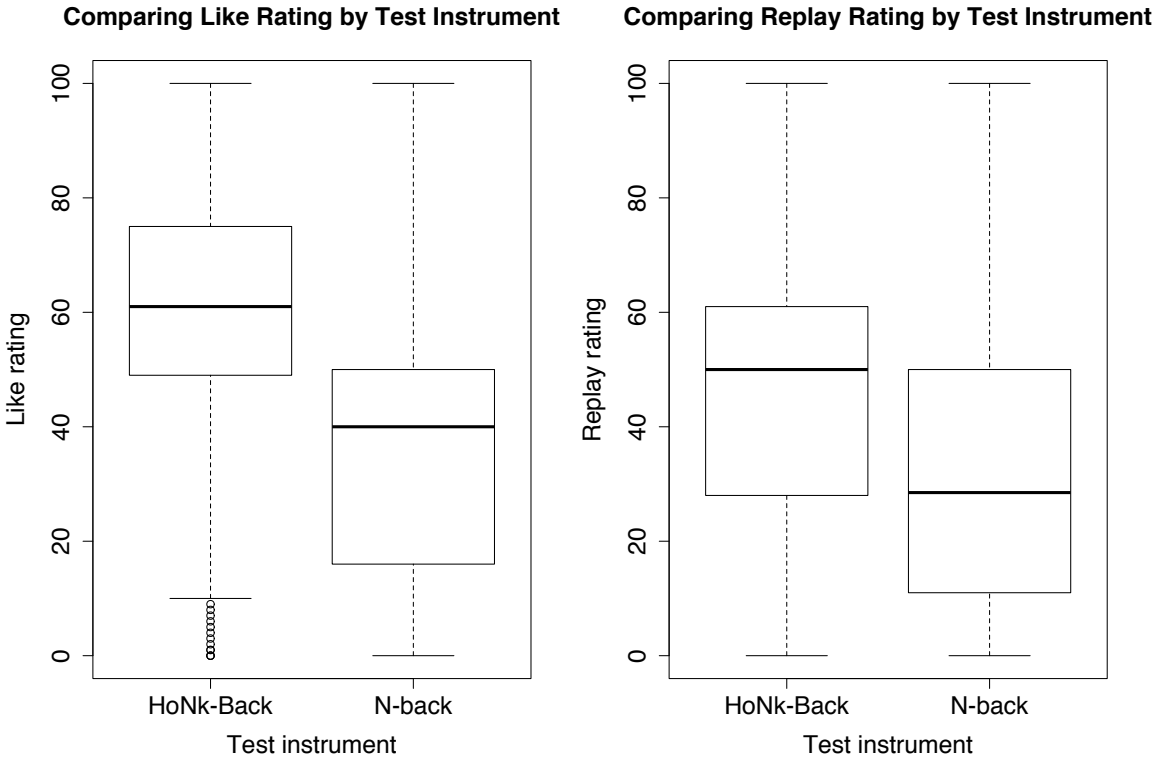
for the final data analysis with an age range from 17 to 73 years ($M = 25.74$, $SD = 14.84$, 3 missings). Participants were randomly assigned to three conditions, 0-back, 1-back and 2-back. They had a twice as likely chance to be assigned to the 1-back or 2-back condition compared to the 0-back condition, resulting in N=46 individuals in the 0-back condition, N=134 in the 1-back and N=104 in the 2-back.

Test instructions were presented in either English or German based on the default browser language, with the option to manually change the language setting. Participants were initially informed that the presented test focuses on attention and working memory. They had to accept the terms presented in the informed consent webpage in order to initiate the testing process. Participants were randomly assigned to start with either the N-back or the HoNk-Back task and were given written instructions for the upcoming tasks. Both the HoNk-Back and N-back had non-skippable trial runs before the main experiment could be initiated. Additionally, the HoNk-Back task instructions were complemented by video tutorials.

3.2.3 Results

Participants liked the HoNk-Back more ($M = 58.1$, $SD = 23.2$) than the N-back ($M = 36.3$, $SD = 22.2$), $t(515) = 10.9$, $p = 2.2\times10^{-16}$, $d = .96$, and were more likely to play the HoNk-Back again ($M = 45.7$, $SD = 23.8$) than the N-back ($M = 31.1$, $SD = 22.5$), $t(462) = 6.84$, $p = 1.29\times10^{-11}$, $d = .63$ (see figure 4). Ratings ranged from 0 (*Not at all*) to 100 (*Very Much*). As of it's nature, the HoNk-Back was rated more

often to be similar to every days activities (N=237) than the N-back (N=16).



(a) (b)

Figure 4. Box plot showing the like (a) and replay (b) rating for the HoNk-Back and the N-back, ranging from 0 (*Not at all*) to 100 (*Very much*). (a) Participants like the HoNk-Back significantly more ($M = 58.1$, $SD = 23.2$) than the N-back ($M = 36.3$, $SD = 22.2$), $t(515) = 10.9$, $p = 2.2{\times}10^{-16}$, $d = .96$. (b) Participants were more likely to replay the HoNk-Back ($M = 45.7$, $SD = 23.8$) than the N-back ($M = 31.1$, $SD = 22.5$), $t(462) = 6.84$, $p = 1.29{\times}10^{-11}$, $d = .63$)

A significant correlation between the HoNk-Back accuracy and N-back accuracy was found (Spearman's $\rho(284) = .50$, $p = 2.2{\times}10^{-16}$). If controlled for the back condition, only the 1-back (Spearman's $\rho(134) = .26$, $p = .002$) and 2-back (Spearman's $\rho(104) = .30$, $p = .002$) showed a significant correlation (0-back (Spearman's $\rho(46) = .28$, $p = .05$)). An upper-bound ceiling effect for the N-back accuracy was observed

for all conditions (see figure 5). This effect is weaker for the HoNk-Back, especially in the 1-back and 2-back condition.

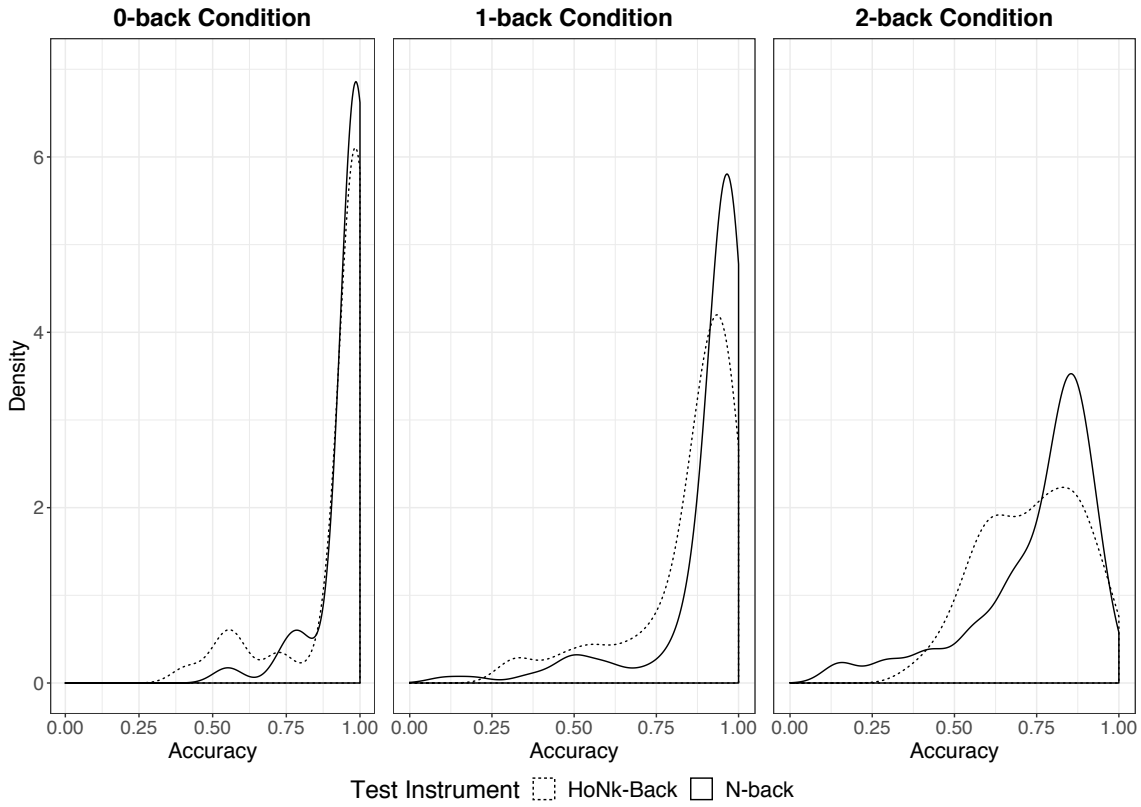


Figure 5. Density plot showing the accuracy of the N-back and HoNk-Back in relation to the test condition. An upper-bound ceiling effect for both the N-back and the HoNk-back can be observed under all conditions, whereas a weaker ceiling effect for the 1-back and 2-back can be observed for the HoNk-Back. N=46, Bandwidth=0.05, N=134, Bandwidth=0.05 and N=104, Bandwidth=0.05 for the 0-back, 1-back and the 2-back condition.

The amount of time spent playing computer games per week by the participants correlates with their accuracy in the HoNk-Back game (Spearman's $\rho(282) = .20$, $p = .001$) but not with the N-back accuracy (Spearman's $\rho(282) = .09$, $p = 0.12$). In addition, participants playing computer games regularly (>2.9 hours a week) have a

significant higher accuracy in the HoNk-Back (M= 0.88, SD = 0.12) than participants not playing regularly (<2.9 hours a week; M = 0.80, SD = 0.18), $t(156)$ = 4.39, p = 1.02×10^{-5}). On contrast, the time spent in a car per week does not correlate with either games accuracy (N-back: Spearman's $\rho(279)$ = .04, p = .47; HoNk-Back: Spearman's $\rho(279)$ = .01, p = .82).

A significant gender effect in the N-back accuracy was found between females (M = 0.83, SD = 0.20) and males (M = 0.90, SD = 0.12), t(146) = −2.84, p = .003, and on the HoNk-Back accuracy as well (females: M = 0.78, SD = 0.18; males: M = 0.89, SD = 0.13), $t(143)$ = −3.75, p = .0001.

4 Discussion

The looming problem of low-powered studies in psychological science still taints its credibility. In this thesis, an article introducing a new concept for online-based large-scale phenotyping and a pilot study examining the early steps of this concept on the basis of a platform hosting the digital gamified testing instruments were presented. With this concept, we call out and face the root of this looming problem - low sample size. For us, the only way to fight this ongoing replication crisis is to power-up the sample sizes of studies. Fortunately, with novel approaches like gamification, applied games and automatization we were able to break the chains of dreariness from classical tests and motivate the user to partake in a *game* rather than a *test*. We showed that study participants significantly like the HoNk-Back more than the

online adaption of the classical N-back and I'd like to emphasize the strong effect size of $d = .96$. Further, the willingness to redo the game was significantly higher than redoing the test, again with a strong effect size of $d = .63$. That is a sign for us that we are taking the right approach with the use of gamification and applied games and successfully implemented the motivational aspects described in the subsection Motivational Aspects. The observed ceiling effect in the 0-back and 1-back conditions is a common result other studies faced, too, when testing a 0-back or 1-back condition (e.g. López-Vicente et al., 2016; Mulquiney, Hoy, Daskalakis, & Fitzgerald, 2011). The rather low correlations for the HoNk-Back and the N-back ($\rho(134) = .26$ and $\rho(104) = .30$ for the 1-back and 2-back, respectively) should be treated with caution as they not fulfill the minimum sample size requirement suggested by Bonett and Wright (2000) and might be the result of said ceiling effect. However, one could cautiously state that these correlations point in the right direction so that the results of the HoNk-Back can be compared and reproduced with those of the N-back, which in turn can be interpreted as a successful translation of a test into an applied game. It is important to note that one of the most frequently mentioned problems that participants had were the controls and interface of the HoNk-Back. The increased difficulty of the 2-back combined with the difficulties to handle the controls and the interface of the HoNk-Back could be one of the explanations why there is no ceiling effect for the HoNk-Back but for the N-back. One could further argue that the controls were almost the same (pressing F and J) for the HoNk-Back and N-back, leaving the interface of

the HoNk-Back as the main source of the problem. Another explanation could be the better discriminatory power that results from the higher ecological validity from the HoNk-Back. Additionally, there was no clear call-to-action in the HoNk-Back, and the participants had trouble playing the game without first receiving an instruction. In a pretest prior to the pilot study, no video instructions were given for the HoNk-Back, resulting in many participants with a 0% to 50% accuracy for the 0-back condition in the HoNk-Back, suggesting problems with the correct use of the interface and the controls. Thus, the game was difficult to learn, less absorbing and less fun, which leads to a high initial cost and a deterrent to potential new participants for playing, resulting in overall fewer participants. In the future development of applied games, one has to make sure that the game can be played without first reading a manual or completing a tutorial. This could mean that more participants play the game and like it even more, which in turn leads to more regular players and more data points.

To sum up, these above mentioned indications point in the right direction, but ironically a larger sample size is necessary to make the concept robust and to eliminate all statistical doubts. Nevertheless, with a redesign of the HoNk-Back interface and more intuitive controls it should be possible to solve the problems and make it easier for new participants to get into the game and experience fun, resulting in a larger sample size. At the end, the set goals for the test battery were only partially met: (a) when above mentioned problems are solved, the screening of subjects for large-scale neurocognitive genetics studies to identify phenotypic extremes should be possible

and feasible; (b) the assessment of parts of human cognition with evidence-based tools was partially met with the results from the pilot study, where the results from the HoNk-Back were comparable with those of the N-back; (c) the HoNk-Back was rated more enjoyable and more likely to be replayed than the N-back, both with high effect sizes (d=.96 and d=.63), thus meeting the goal of an entertaining and enjoyable user experience completely.

The ethical aspect plays a central role when personal data is involved. In this context, data security should always be a major issue, as the article by De Montjoye, Radaelli, Singh, et al. (2015) well illustrates, where anonymised metadata of credit card records were analyzed and only four spatiotemporal points were necessary to reidentify 90% of individuals. Anonymization by removing obvious identifiers like name, home address or telephone number, as suggested under the U.S. personally identifiable information approach, does not make the metadata completely anonymous and should be considered unsafe to release to third-parties or the public (De Montjoye et al., 2015). This can also be generalized to other data sets, such as metadata of telecommunication companies, where spatiotemporal points are also tracked and stored. If such metadata were to fall into the wrong hands, it could be exploited in many illegal ways, as knowledge of a person's whereabouts at any given time is highly delicate information. Hence, the selling of data like in the Facebook example mentioned in the section Data Privacy & Security can be considered highly unethical. Dealing with personal data of this scope should be strictly regulated, also with regard

to the anonymization of the dataset. Why anonymization of datasets is important highlights the data leak in 2015 of the dating site *Ashley Madison*, a website that enables extramarital affairs. There, personal data of 37 million users was released online after a hack (Mansfield-Devine, 2015). After this data leak, it was reported on multiple instances that people committed suicide because their name and personal information appeared publicly available online in the leaked data (Mansfield-Devine, 2015).

The ethical aspect grows even further as soon as medical patient data in the form of a electronic patient records are involved. There are numerous reports about data breaches in hospitals and other medical institutions, as for example in a report by Gabriel, Noblin, Rutherford, Walden, and Cortelyou-Ward (2018), where they identified that between 2009 and 2016 data breaches in hospitals accounted for about 30% of large data security incidents that were reported to the U.S. Department of Health and Human Services Office for Civil Rights. They further state that security breaches continue to affect hospitals even with sophisticated IT infrastructure (Gabriel et al., 2018) and Pahnila, Siponen, and Mahmood (2007) summarizes that the biggest threat to information systems are employees who do not adhere to the security protocols. Therefore, it is sadly not shocking that a large health data leak was just recently reported in the news, where millions of patient dossiers and pictures from patients in 50 countries, including Germany and Switzerland, were freely accessible on the Internet (Maximilian Zierer, 2019). Such news, of course, rightly spurs on the current

debate on patient data protection even more. This will give more attention to the topic and hopefully such large data leaks will no longer occur on such a large scale or not at all in the future. However, this can only happen if employees and the public are educated and made aware of the data protection concept and how to deal with it properly.

During the writing of this thesis, the Bundestag in Germany adopted a law (*Digitale-Versorgung-Gesetz*, DVG, (Deutscher Bundestag, 2019)) on November 7th that states that as of 2021, laboratory findings, diagnoses and treatment data are to be entered in a digital patient file, digitally accessible to the patient. This law is intended to provide a great opportunity for better health care in Germany, as the opening sentence in the law states (Deutscher Bundestag, 2019). Even if this sounds appealing in theory, there is much criticism from data protectors concerning this law (for e.g. Waschinski, 2019). Against this criticism, the state-owned Danish platform *sundhed.dk* shows how well such a system, where patients can access their patient dossier online, can work. There, 5.8 million inhabitants, doctors and pharmacists have access to patient data and around 1.7 million hits per day are registered (Wünnenberg, 2017), which shows that it is regularly used. The future will show whether the digitalization of patient dossiers can be successfully implemented, not just in Germany, but worldwide, and whether data security and data privacy can be mastered, because health data digitalization is extremely problematic without it.

In summary, the ethical aspect connected to personal data and patient data must

not be ignored or taken lightly. It is essential in the near future to assure strict laws and guidelines towards data security, especially and above all in the digital health sector, where patient data is extraordinarily sensitive and valuable. Should the health data be linked with other data (e.g. search history, browsing behavior, credit card transaction information) and fall into the hands of, for example, health insurance companies, it is plausible that the entire health insurance system will be revolutionized. A first step in this direction is being made with the use of health monitors such as *FitBit* (`www.fitbit.com`), where insurance companies give their customers discounts or other benefits if the customer makes the data from the device accessible to the insurance company (e.g. CSS Versicherungen, Helsana). This raises many ethical questions and will therefore remain an important issue for a long time to come.

The article presented in the scope of this thesis highlights our established platform COSMOS which contains a digital psychometric toolkit that enables automatized psychometric data collection while measuring a broad range of cognitive functions. We use the approach of applied games and gamification to overcome some of the most pressing problems psychological science faces as it allows a low cost individual psychometric testing and scalable assessment while putting the user in the foreground by focusing on an entertaining experience. That we succeeded in some respects is shown by the results of the pilot study presented. There, we showed that participants like the applied game HoNk-Back more than the non-gamified N-back and are more

likely to replay the HoNk-Back, both showing a strong effect. However, we did not achieve the sample size we hoped for, which shows us that the provided intrinsic motivation for participate in the test was still too small. Another reason could be the steep learning curve that the HoNk-Back has: some participants reported that they did not understand the interface and the controls and that might lead to frustration and resentment. Furthermore, the provided custom feedback could simply be not enough incentive to spend 15 minutes of one's time on a online test. While we could not motivate participants to partake totally voluntarily, the scalability and low cost of individual testing for screening large cohorts could save substantial expenses by not having to pay for a physical laboratory, testing computers, test program licenses or an investigator but only a small remuneration to increase motivation.

To sum up, I have purposed a platform for large-scale phenotyping with applied games that has the potential to overcome the replication crisis by fighting it's core problem: low sample sizes. That goal was only partially reached and the provided motivation was not enough for participants to partake in large numbers in our pilot study. However, the gamification of a widely used performance task in cognitive neuroscience, the N-back, was successful in regards of higher appeal and replayability. Furthermore, I showed a way for very efficient screening of neurocognitive phenotypes that can help future large-scale studies cut costs, allowing for a potential larger sample size. On this basis, future studies can use the in this thesis outlined benefits of applied games in the automated online collection of neurocognitive phenotypes and cohort

screenings. Personal and health data collected in this way and how to deal with them is and will remain a delicate matter and will fuel many more discussions. If we overcome the delicate matter of data privacy and data security connected to personal and health data, we can make a step closer to end the replication crisis that is tainting various sciences for more than five decades and then might soon become a thing of the past.

5 References

Ahmed, A. A., & Khay, L. M. (2017). Securing user credentials in web browser: review and suggestion. In *2017 ieee conference on big data and analytics (icbda)* (pp. 67–71).

Anderson, D., & Hills, M. (2017). Query construction patterns in php. In *2017 ieee 24th international conference on software analysis, evolution and reengineering (saner)* (pp. 452–456).

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547-1562. doi: 10.1177/0956797617723724

Bauman, E., Lu, Y., & Lin, Z. (2015). Half a century of practice: Who is still storing plaintext passwords? In *International conference on information security practice and experience* (pp. 253–267).

Begley, C. G., & Ellis, L. M. (2012, 03 28). Raise standards for preclinical cancer research. *Nature*, *483*, 531 EP -.

Bonett, D. G., & Wright, T. A. (2000, Mar 01). Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, *65*(1), 23–28. doi: 10 .1007/BF02294183

Breur, T. (2016, Jul 01). Statistical power analysis and the contemporary "crisis" in social sciences. *Journal of Marketing Analytics*, *4*(2), 61–65. doi: 10.1057/s41270-016-0001 -3

Browne, K., & Anand, C. (2012). An empirical evaluation of user interfaces for a mobile video game. *Entertainment Computing*, *3*(1), 1–10.

Brüne, M. (2005, 01). "Theory of Mind" in Schizophrenia: A Review of the Literature. *Schizophrenia Bulletin*, *31*(1), 21-42. doi: 10.1093/schbul/sbi002

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. M.-a., Dawson, D. R., ... Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society : JINS*, *12*(2), 194–209. doi: 10.1017/S1355617706060310

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *Sat*, *17*, 22–03.

Colzato, L. S., Van Wouwe, N. C., Lavender, T. J., & Hommel, B. (2006). Intelligence and cognitive flexibility: fluid intelligence correlates with feature "unbinding" across perception and action. *Psychonomic Bulletin & Review*, *13*(6), 1043–1048.

Cox, J. (2016). *The administrator of the dark web's infamous hacking market has vanished.*

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450–466.

Debnath, S., Chattopadhyay, A., & Dutta, S. (2017). Brief review on journey of secured hash algorithms. In *2017 4th international conference on opto-electronics and applied optics (optronix)* (pp. 1–5).

Deepa, G., & Thilagam, P. S. (2016). Securing web applications from injection and logic vulnerabilities: Approaches and challenges. *Information and Software Technology*, *74*, 160 - 180. doi: 10.1016/j.infsof.2016.02.005

De Montjoye, Y.-A., Radaelli, L., Singh, V. K., et al. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539.

Deterding, S. (2012). Gamification: designing for motivation. *interactions*, *19*(4), 14–17.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic mindtrek conference: Envisioning future media environments* (pp. 9–15).

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In *Chi '11 extended abstracts on human factors in computing systems* (pp. 2425–2428). New York, NY, USA: ACM. doi: 10.1145/1979742.1979575

Deutscher Bundestag. (2019). *Entwurf eines gesetzes für eine bessere versorgung durch digitalisierung und innovation (digitale-versorgung-gesetz – dvg, 19/13438).*

Di Giacomo, M. (2005). Mysql: lessons learned on a digital library. *IEEE software*, *22*(3), 10–13.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012, 01). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE*, *7*(1), 1-7. doi: 10.1371/journal.pone.0029081

Durumeric, Z., & Kasten, J. (2013). Analysis of the HTTPS certificate ecosystem. *Proceedings of the 2013 conference on Internet measurement conference*, 291–304. doi: 10.1145/2504730.2504755

Eich, E. (2014). *Business not as usual.* Sage Publications Sage CA: Los Angeles, CA.

Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 871–880).

Elms, A. C. (1975). The crisis of confidence in social psychology. *American psychologist*, *30*(10), 967.

Emond, M. J., Louie, T., Emerson, J., Zhao, W., Mathias, R. A., Knowles, M. R., ... GO, L. (2012). Exome sequencing of extreme phenotypes identifies dctn4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis. *Nature Genetics*, *44*(8), 886–889. doi: 10.1038/ng.2344

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PloS one*, *11*(2).

European Parliament. (2016, April). *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (data protection directive).*

Fleming, T. M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S. N., Cheek, C., ... Hetrick, S. (2017). Serious games and gamification for mental health: Current status and promising directions. *Frontiers in Psychiatry*, *7*, 215. doi: 10.3389/fpsyt.2016.00215

Fleming, T. M., Cheek, C., Merry, S. N., Thabrew, H., Bridgman, H., Stasiak, K., ... Hetrick, S. (2014). Serious games for the treatment or prevention of depression: a systematic review.

Francisco-Aparicio, A., Gutiérrez-Vela, F. L., Isla-Montes, J. L., & Sanchez, J. L. G. (2013). Gamification: Analysis and application. In V. M. Penichet, A. Peñalver, & J. A. Gal-

lud (Eds.), *New trends in interaction, virtual reality and modeling* (pp. 113–126). London: Springer London. doi: 10.1007/978-1-4471-5445-7_9

Frost, S. (2015). *Drowning in big data? reducing information technology complexities and costs for healthcare organizations.*

Gabriel, M. H., Noblin, A., Rutherford, A., Walden, A., & Cortelyou-Ward, K. (2018). Data breach locations, types, and associated characteristics among us hospitals. *Am J Manag Care*, *24*(2), 78–84.

Gartner. (2019, October). *Big data.* Retrieved 2019-10-03, from `https://www.gartner.com/it-glossary/big-data/`

Goseva-Popstojanova, K., Anastasovski, G., Dimitrijevikj, A., Pantev, R., & Miller, B. (2014). Characterization and classification of malicious web traffic. *Computers & Security*, *42*, 92 - 115. doi: 10.1016/j.cose.2014.01.006

Graafland, M., Dankbaar, M., Mert, A., Lagro, J., De Wit-Zuurendonk, L., Schuit, S., ... Schijven, M. (2014, Nov 11). How to systematically assess serious games applied to health care. *JMIR Serious Games*, *2*(2), e11. doi: 10.2196/games.3825

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications.* Allyn & Bacon.

Groh, F. (2012). Gamification: State of the art definition and utilization. *Institute of Media Informatics Ulm University*, *39*, 31.

Guin, T. D.-L., Baker, R., Mechling, J., & Ruyle, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research*, *54*(5), 613–633.

Hamari, J., Koivisto, J., Sarsa, H., et al. (2014). Does gamification work?-a literature

review of empirical studies on gamification. In *Hicss* (Vol. 14, pp. 3025–3034).

Harrington, L., Siegert, R., & McClure, J. (2005). Theory of mind in schizophrenia: A critical review. *Cognitive Neuropsychiatry*, *10*(4), 249-286. doi: 10.1080/13546800444000056

Heitz, R. P., Unsworth, N., & Engle, R. W. (2005). Working memory capacity, attention control, and fluid intelligence. *Handbook of understanding and measuring intelligence*, 61–77.

Idrees, S. M., Alam, M. A., & Agarwal, P. (2018). A study of big data and its challenges. *International Journal of Information Technology*, 1–6.

Ioannidis, J. P. A. (2005, 08). Why most published research findings are false. *PLOS Medicine*, *2*(8). doi: 10.1371/journal.pmed.0020124

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, affective & behavioral neuroscience*, *7*(2), 75–89. doi: 10.3758/CABN.7.2.75

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N -back task as a working memory measure. *Memory*, *18*(4), 394–412. doi: 10.1080/09658211003702171

Kidd, G. R., & Humes, L. E. (2015). Keeping track of who said what: Performance on a modified auditory n-back task with young and older adults. *Frontiers in Psychology*, *6*(July), 1–15. doi: 10.3389/fpsyg.2015.00987

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Kosara, R., & Haroz, S. (2018). Skipping the replication crisis in visualization: Threats to study validity and how to address them: Position paper. In *2018 ieee evaluation and beyond-methodological approaches for visualization (beliv)* (pp. 102–107).

Krawczyk, H., Paterson, K. G., & Wee, H. (2013). On the security of the tls protocol: A systematic analysis. In *Annual cryptology conference* (pp. 429–448).

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming*, *45*(6), 752–768.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660-664. doi: 10.1177/1745691616687745

Limperos, A. M., Schmierbach, M. G., Kegerise, A. D., & Dardis, F. E. (2011). Gaming across different consoles: Exploring the influence of control scheme on game-player enjoyment. *Cyberpsychology, Behavior, and Social Networking*, *14*(6), 345-350. doi: 10.1089/cyber.2010.0146

López-Vicente, M., Forns, J., Suades-González, E., Esnaola, M., García-Esteban, R., Álvarez-Pedrerol, M., ... Sunyer, J. (2016). Developmental trajectories in primary schoolchildren using n-back task. *Frontiers in Psychology*, *7*, 716. doi: 10.3389/fpsyg.2016.00716

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. *JMIR Serious Games*, *4*(2), e11. doi: 10.2196/games.5888

Malvoni, K., & Knezovic, J. (2014). Are Your Passwords Safe: Energy-Efficient Bcrypt

Cracking with Low-Cost Parallel Hardware. In *Proceedings of the 8th usenix workshop on offensive technologies - woot '14* (p. 10).

Mansfield-Devine, S. (2015). The ashley madison affair. *Network Security*, *2015*(9), 8–16.

Maximilian Zierer, H. T. (2019, September). *Millionenfach patientendaten ungeschützt im netz.* Retrieved 2019-11-10, from `https://www.br.de/nachrichten/deutschland-welt/millionenfach-patientendaten-ungeschuetzt-im-netz`

McAfee, I. (2014). *Net losses: Estimating the global cost of cybercrime: Economic impact of cybercrime II* (Tech. Rep. No. June). Center for Strategic and Int'l Studies. Retrieved from `www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime2.pdf`

Mirza-Babaei, P., Nacke, L. E., Gregory, J., Collins, N., & Fitzpatrick, G. (2013). How does it play better?: Exploring user testing and biometric storyboards in games user research. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1499–1508). New York, NY, USA: ACM. doi: 10.1145/2470654.2466200

Moore, D., & Rid, T. (2016). Cryptopolitik and the darknet. *Survival*, *58*(1), 7-38. doi: 10.1080/00396338.2016.1142085

Mulquiney, P. G., Hoy, K. E., Daskalakis, Z. J., & Fitzgerald, P. B. (2011). Improving working memory: Exploring the effect of transcranial random noise stimulation and transcranial direct current stimulation on the dorsolateral prefrontal cortex. *Clinical Neurophysiology*, *122*(12), 2384 - 2389. doi: 10.1016/j.clinph.2011.05.009

Muscat, I. (2016). Web vulnerabilities: identifying patterns and remedies. *Network Security*, *2016*(2), 5 - 10. doi: 10.1016/S1353-4858(16)30016-2

Ninaus, M., Pereira, G., Stefitz, R., Prada, R., Paiva, A., Neuper, C., & Wood, G. (2015).

Game elements improve performance in a working memory training task. *International Journal of Serious Games*, *2*(1), 3–16. doi: 10.17083/ijsg.v2i1.60

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... others (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137-141. doi: 10.1027/1864-9335/a000192

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., ... Shakarian, P. (2016). Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *2016 ieee conference on intelligence and security informatics (isi)* (pp. 7–12).

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. In *Human brain mapping* (Vol. 25, pp. 46–59). doi: 10.1002/hbm.20131

Pahnila, S., Siponen, M., & Mahmood, A. (2007). Employees' behavior towards is security policy compliance. In *2007 40th annual hawaii international conference on system sciences (hicss'07)* (pp. 156b–156b).

Pashler, H., Coburn, N., & Harris, C. R. (2012, 08). Priming of social distance? failure to replicate effects on social and food judgments. *PLOS ONE*, *7*(8), 1-6. doi: 10.1371/journal.pone.0042510

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528-530. doi: 10.1177/1745691612465253

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319-332. doi:

10.1177/1745691614528519

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. doi: 10.1017/S0140525X00076512

Procaccianti, G., Fernández, H., & Lago, P. (2016). Empirical evaluation of two best practices for energy-efficient software development. *Journal of Systems and Software*, *117*, 185–198.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102–1113. doi: 10.3758/s13423-013-0453-9

Ritterfeld, U., Cody, M., & Vorderer, P. (2009). *Serious games: Mechanisms and effects.* Routledge.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54–67.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems (cts)* (pp. 42–47).

Schmierbach, M., Chung, M.-Y., Wu, M., & Kim, K. (2014). No one likes to lose. *Journal of Media Psychology*.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, *70*, 263 - 286. doi: 10.1016/j.jbusres.2016.08.001

Spenneberg, R. (2009). Webserver-sicherheit mit mod_security. *Datenschutz und Datensicherheit-DuD*, *33*(3), 155–160.

Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality frame-

work. In *2015 ieee international congress on big data* (pp. 191–198).

Tene, O., & Polonetsky, J. (2011). Privacy in the age of big data: a time for big decisions. *Stan. L. Rev. Online*, *64*, 63.

Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015). A survey on security and privacy issues in big data. In *2015 10th international conference for internet technology and secured transactions (icitst)* (pp. 202–207).

The Economist. (2017, May). *The world's most valuable resource is no longer oil, but data.* Retrieved 2019-09-15, from `https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data`

Tipton, S. J., & Choi, Y. B. (2016). Toward secure web application design: Comparative analysis of major languages and framework choices. *International Journal of Advanced Computer Science and Application*, *7*(2), 48–53.

Trautman, L. J., & Ormerod, P. C. (2016). Corporate directors' and officers' cybersecurity standard of care: The yahoo data breach. *Am. UL Rev.*, *66*, 1231.

Trepte, S., & Reinecke, L. (2011). The pleasures of success: Game-related efficacy experiences as a mediator between player performance and game enjoyment. *Cyberpsychology, Behavior, and Social Networking*, *14*(9), 555–557.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, *28*(2), 127–154.

van den Hoogen, W., Poels, K., IJsselsteijn, W., & de Kort, Y. (2012). Between challenge and defeat: Repeated player-death and game enjoyment. *Media Psychology*, *15*(4), 443-459. doi: 10.1080/15213269.2012.723117

Wang, S., Chen, Y., Jiang, W., Li, P., Dai, T., & Cui, Y. (2009). Fairness and interactivity of

three cpu schedulers in linux. In *2009 15th ieee international conference on embedded and real-time computing systems and applications* (pp. 172–177).

Waschinski, G. (2019, May). *Mängel beim datenschutz“ – opposition kritisiert spahns vorgehen bei digitaler patientenakte.* Retrieved 2019-11-10, from `https://www.handelsblatt.com/24364532.html`

Wiemer, F., & Zimmermann, R. (2014). High-speed implementation of bcrypt password search using special-purpose hardware. *2014 International Conference on Reconfigurable Computing and FPGAs, ReConFig 2014*. doi: 10.1109/ReConFig.2014.7032529

Winn, B., & Heeter, C. (2006). Resolving conflicts in educational game design through playtesting. *Innovate: Journal of Online Education*, *3*(2).

Wired. (2014, July). *Data is the new oil of the digital economy.* Retrieved 2019-09-15, from `https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/`

Wünnenberg, I. (2017, November). *Die hürden der elektronischen patientenakte.* Retrieved 2019-11-10, from `https://heise.de/-3902829`

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. *Ieee Access*, *2*, 1149–1176.

Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: architecture and challenges. *IEEE Network*, *28*(4), 5–13.

6 Declaration by Candidate

I declare herewith that I have independently carried out the PhD-thesis entitled "*Applied Games for smart phenotypic data acquisition - challenges of a web platform with digital gamified testing instruments for online-based large scale phenotyping*". This thesis consists of original research articles that have been written in cooperation with the enlisted co-authors and have been published in peer-reviewed scientific journals or are in preparation for publication / submitted for publication. Only allowed resources were used and all references used were cited accordingly.

Date: _____

Signature: _____

7 Appendix

7.1 Security Aspects

7.1.1 Linux, Apache HTTP Server & MySQL Database

Linux is the most commonly used open source operating system and is very popular in server environments (Wang et al., 2009). Consequently, the Linux distribution *Ubuntu Server* was chosen as the fundamental operating system for COSMOS. Upon Linux, we used an *Apache HTTP Server* and a *MySQL Database Server* as the back-end software, both well-known and widely-used open source software applications (Procaccianti, Fernández, & Lago, 2016).

The Apache HTTP Server (short: *Apache*) is a web server software that serves COSMOS to the public. Apache is considered to be a secure web server software because of it's optional security module "*mod_security*" (Spenneberg, 2009), which we enabled and configured for COSMOS.

MySQL Database Server's (short: *MySQL*) speed, connectivity, scalability, reliability and security (Di Giacomo, 2005) makes it well suited for our use-case. The synergy between MySQL and Apache has made MySQL the first choice as a database for our use-case. Our database consists of 6 separate password-protected application-specific tables, one for each implemented game and one for the users information and the platforms metadata. Game tables have read-only restrictions to the COSMOS table User-ID field, which is a unique identifier code (UIC) for the user, and can not

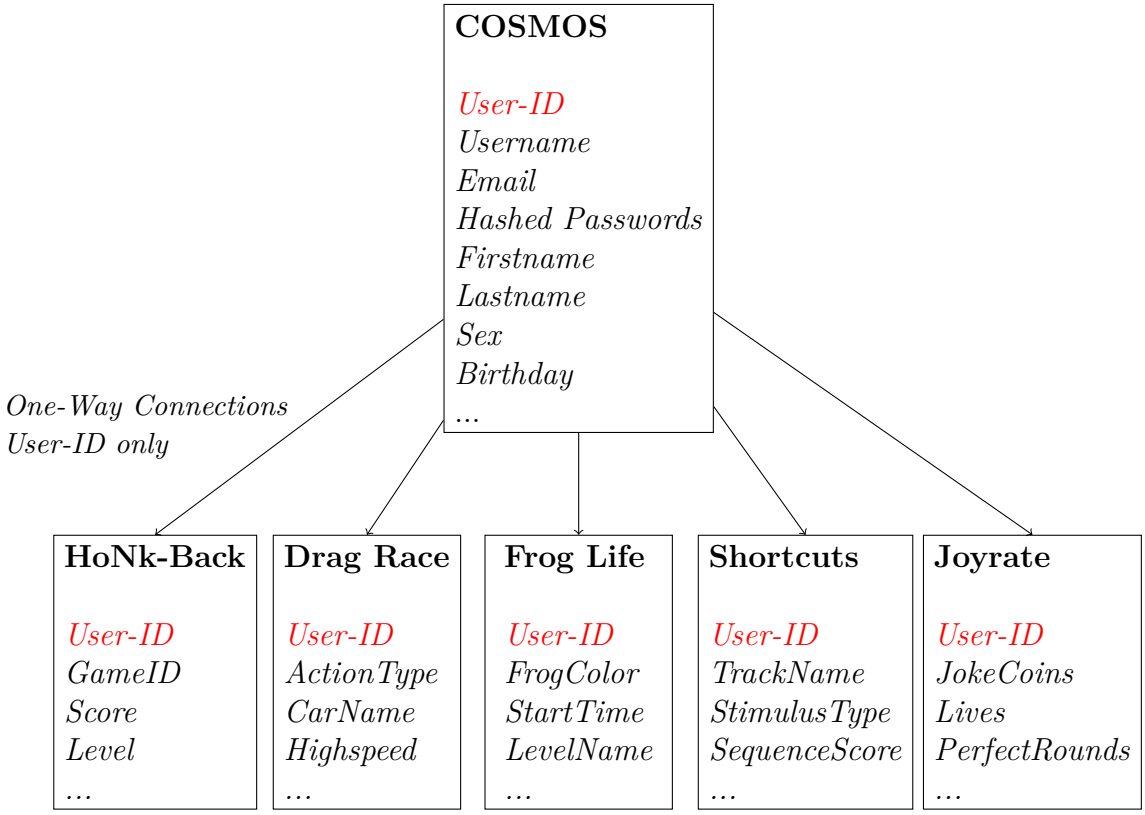


Figure 6. Schematic figure of the six digitally separated tables of COSMOS and the five games. As soon as a game is started by a registered user it reads the users User-ID from the COSMOS table and utilizes it for the identification of the generated data. The User-ID, a place holder for the users information, is transferred in a one-way connection to the game tables. The game tables can not communicated horizontally or back to COSMOS, boosting security.

communicate with each other, rendering them isolated, making the database structure safer. However, should a user decide not to register with COSMOS, the User-ID will be replaced by a randomly generated Session-ID, unique to to this particular user, preventing anonymous data generation by potential misuse of the tools. This increases security not only for the users information, but also for the users generated data. Figure 6 illustrates this schema.

7.1.2 Zend Framework 2

The *Zend Framework 2* (short: *Zend*) was used as a programming framework. It is a web application framework and offers the most robust set of tools to achieve security (Tipton & Choi, 2016). Zend is based on the programming language *PHP*, a programming language that unites three major advantages: 1. it is compatible with Linux, 2. its high performance with MySQL queries (Procaccianti et al., 2016) and 3. it is one of the most popular programming language for websites (D. Anderson & Hills, 2017) with a large community.

One of Zends major strength are the out-of-the-box security components, many of which COSMOS utilizes, as e.g. CSRF is used by default for forms; input sanitation preventing database injections; storing passwords more efficiently and securely and the encrypted authentication (Tipton & Choi, 2016). Lastly, there are several ways to securely store a password, such as password hashing.

7.1.3 Password Hashing

Password hashing is a subcategory of cryptography (Debnath, Chattopadhyay, & Dutta, 2017) and is a technique where passwords are converted from plain text into an unrecognizable sequence of characters (Ahmed & Khay, 2017). It has been shown that password hashing is the correct method for storing passwords, provided that it is used with a suitable hashing algorithm like *bcrypt* (Bauman, Lu, & Lin, 2015). bcrypt is widely used as default password hashing algorithm in common program-

ming languages, such as PHP (Wiemer & Zimmermann, 2014) and was designed to be secure for several years, even with improving hardware (Malvoni & Knezovic, 2014). COSMOS hashes the users password by using bcrypt before storing it into the database.

Additionally to this technical precaution, when registering for an account at COSMOS the user is provided with a real-time while-typing password strength meter that indicates whether the chosen password is considered weak or strong, thus further encouraging the user to choose a strong password.

That being said, hashing passwords does not completely protect these passwords from being translated into plain text. The major advantage of hashing passwords with such a strong algorithm like bcrypt is its immensely cost for a cyber criminal to decipher a stolen hashed password, render it unprofitable to even consider stealing those passwords. For example, to crack a 8-character password (containing letters, numbers and special characters) hashed with bcrypt in one month, one needs hardware that costs approximately between 6 and 9 million dollars (Wiemer & Zimmermann, 2014).

However, hashing a password at the end of the connection is useless if the connection itself and the transmitted data is not encrypted. That's when cryptographic protocols that encrypt the connection come into play.

7.1.4 TLS

The cryptographic protocol *TLS* is the most commonly used protocol for securing communications on the Internet (e.g. between user and a server (Durumeric & Kasten, 2013)) and is considered to be the most important real-world application of cryptography (Krawczyk, Paterson, & Wee, 2013). We used TLS [2] to encrypt and secure the connection and data between the user and COSMOS. See figure 7 for an illustration of the secure connection setup.

7.1.5 Proxy Server

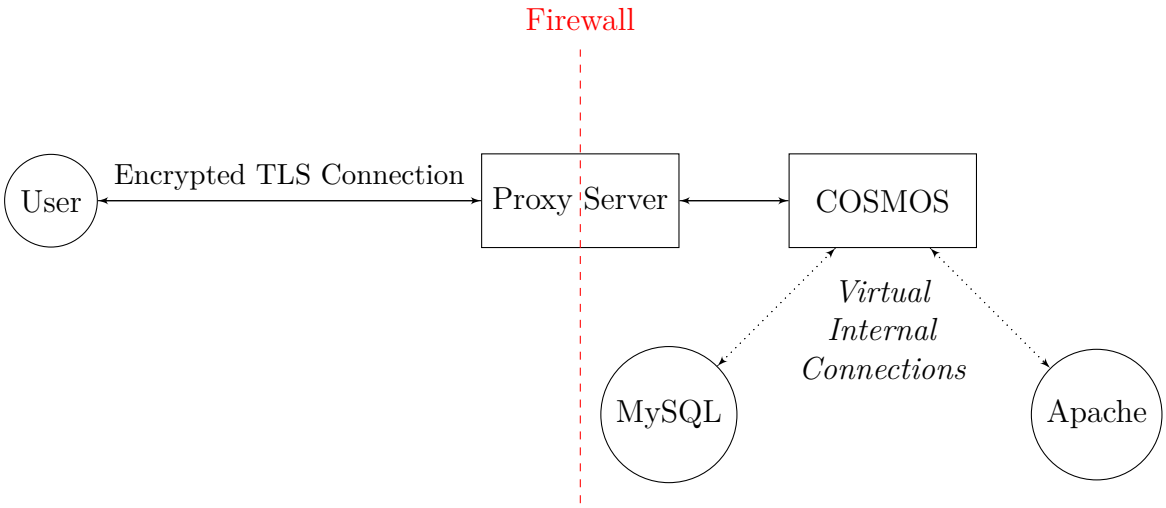


Figure 7. Schematic figure of the hardware setup. The user accesses the COSMOS URL and connects to the COSMOS server. The connection runs via a Proxy Server, which acts as an intermediary and checks the received data for malicious content, only forwarding the data when no threats were found. This connection is encrypted using the cryptographic protocol TLS.

A *Proxy Server* (short: *Proxy*) functions as an intermediary server for user re-

[2] Exact version: TLS ECDHE RSA WITH AES 128 GCM SHA256, 128-bit key, TLS 1.2

quests, thus preventing direct access to the COSMOS server. Furthermore, the proxy scans all incoming data packets for malicious code and only passes the data to the COSMOS server if no threats in the code were found, acting as a firewall. However, the scanned data packets are only stored temporarily for a couple of seconds, deeming no threat to the users privacy. On the other hand, the COSMOS server communicates directly and solely with the Proxy and is unaware of a direct connection to the user. Figuratively speaking, the Proxy acts as a trustee between the users and the COSMOS server, increasing the security for both parties. To even further increase the security, the COSMOS server has been set to exclusively accept direct data connections from the proxy and denying all other sources. Figure 7 illustrates the server setup including the Proxy.

7.2 COSMOS Pilot Study

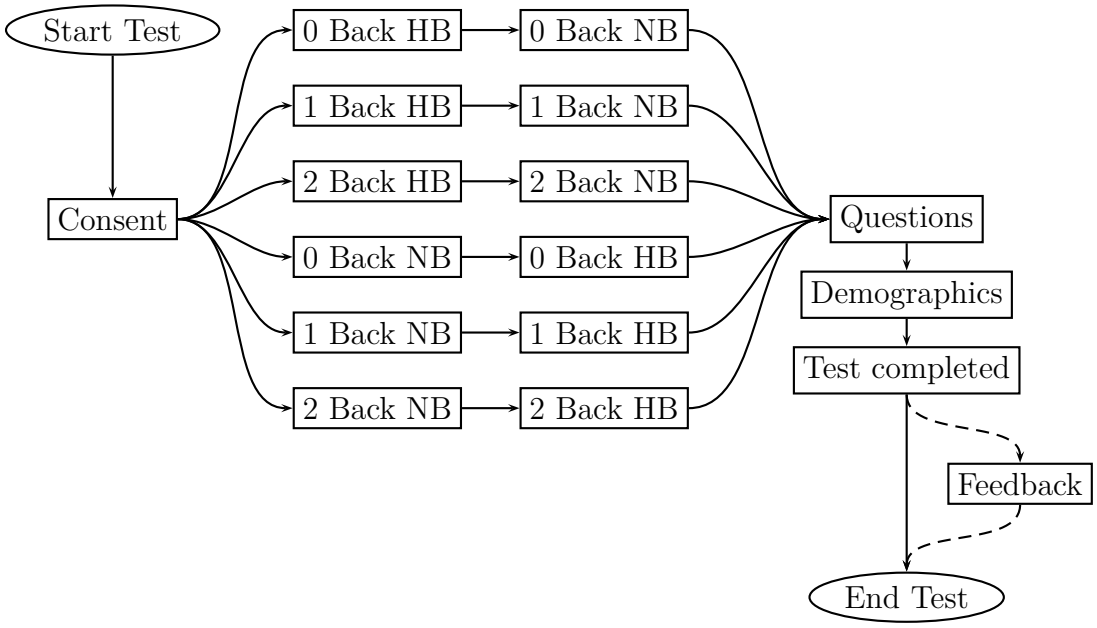


Figure 8. Schematic figure of the test setting in the COSMOS pilot study. After the participants agreed to the consent they were randomly assigned to start with either the N-back (*NB*) or the HoNk-Back (*HB*) task. Further, they were randomly assigned to one of the three conditions: 0-back, 1-back or 2-back, whereas the 1-back and 2-back have twice the chance of being selected. This resulted in N=46 individuals in the 0-back condition, N=134 in the 1-back and N=104 in the 2-back. After completion of the tasks, participants were redirected to the questionnaire section including demographics, rating the appeal of and readiness to repeat the completed tasks, weekly hours of engaging in computer games, weekly driving hours in a car and an open-feedback field. At the end, the participants could choose whether they wanted to have automatically generated feedback on their performance, which would allow them to compare themselves with the previous participants of the experiment.